Big Data

云端时代杀手级应用
大数据分析

胡世忠 著

源源不绝的**庞杂数据量**，**彻底改变**游戏规则，
谁能**理出脉络**、**洞察商机**、**领先创新**，就能成为**新赢家**！









云端时代杀手级应用
大数据分析

胡世忠　著

人民邮电出版社
北京

图书在版编目（CIP）数据

云端时代杀手级应用 ：大数据分析 / 胡世忠著. —
北京 ：人民邮电出版社，2013.6
ISBN 978-7-115-31991-3

Ⅰ. ①云… Ⅱ. ①胡… Ⅲ. ①统计数据—统计分析—
研究 Ⅳ. ①O212.1②F222.1

中国版本图书馆CIP数据核字（2013）第102706号

内 容 提 要

本书分什么是大数据、大数据大商机、技术与前瞻 3 个部分。第一部分介绍大数据分析的概念，以及企业、政府部门可应用的范畴。什么是大数据分析？与个人与企业有什么关系？将对全球产业造成怎样的冲击？第二部分完整介绍大数据在各产业的应用实况，为企业及政府部门提供应用的方向。提供了全球各地的实际应用案例，涵盖零售、金融、政府部门、能源、制造、娱乐、医疗、电信等各个行业，充分展现大数据分析产生的效益。第三部分则简单介绍了大数据分析所需技术及未来发展趋势，为读者提供了应用与研究的方向。

本书以科普的视角描述了云端时代的大数据分析应用，涉及各个行业，具有普遍的参考性和指导性，适合所有对数据、数据挖掘、数据分析感兴趣的技术人员、管理人员、咨询人员、企业决策者及相关行业人员阅读。

◆ 著　　　　胡世忠
责任编辑　杜　洁
责任印制　程彦红　焦志炜

◆ 人民邮电出版社出版发行　　北京市崇文区夕照寺街 14 号
邮编 100061　　电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京铭成印刷有限公司印刷

◆ 开本：720×960 1/16
印张：16
字数：205 千字　　　　2013 年 6 月第 1 版
印数：1-4 000 册　　　2013 年 6 月北京第 1 次印刷

定价：45.00 元

读者服务热线：(010)67132692　印装质量热线：(010)67129223
反盗版热线：(010)67171154

推荐序一

在数据海中航向创新之地

每一次的技术革命都意味着外部环境的大幅改变，而企业也必须跟着转型，否则很快就会在时代的巨轮下消逝。而转型，以 IBM 自身来说，有超过 100 年的经验可以分享。

经过前 50 年的迅速发展，到 20 世纪 70 年代的时候，IBM 已经是在计算机行业里非常成功的公司，但是 80 年代之后，IBM 犯了一系列的错误，在 1993 年陷入最低谷，甚至濒临破产。

市场的翻转，让我们走上一条不断地改变自己、去适应环境变化的转型和创新之路。1993 年之后，IBM 历经了 3 次重大的转型，第 1 次转型是从硬件转向软件和服务，第 2 次转型则是出售了 PC 业务，向高价值业务转型。2004 年之后，整合“全球企业”和“智慧的地球”，让 IBM 每年节约开支 10 亿美元，每股盈利连续 8 年、36 个季度以两位数增长。而今，我们 90%的利润来自软件和服务，而不是来自硬件。

过去 10 年，大环境改变是很大的，尤其在金融危机以后。然而，成功的时代里也有失败的企业，即便是在最困难的环境里也有非常成功的企业，唯有转型和创新，才有能力驾驭这些外来的影响。

2012 年，IBM 在全球范围内对 1709 位企业 CEO 进行了调研，了解了他们对互联经济大格局之下企业如何持续转型的看法，发现 CEO 们并不满足于 IT 运用仅止于整合供应链和后端办公系统，而是希望充分发挥大数据和互联科技的潜力，重新思考人与人互联后对企业带来的价值。

有超过一半的 CEO 认为，在新时代里要转型创新，必须建立广泛的伙伴关系，进行充分的协作。超过 70%的 CEO 认为，企业必须建立强大的业务分析洞察能力，从而能够深入理解每一个客户，对于他们的需求做出快速反应，以个性化服务赢得客户。也有 75%的 CEO 认为，在新的互联时代背景下，社交媒体的影响越来越大，必须有新的人才战略。

这些来自绩优企业的 3 项最重要的发现，代表着企业必须建立更加开放的文化，抓住世界进入计算新纪元的趋势和契机，挖掘出大数据中有价值的洞察，而这些洞察也将成为企业创新的源泉，以及和客户一起智慧增长的能力。

尤其在现今的环境里，数据量的爆发和以前完全不一样，我们的数据有 90%是过去两年创造出来的，到了 2020 年的时候，全世界要消化的数据量是现在的 44 倍以上。为什么有那么大的数据源？除了传统企业掌握的数字，每天都有各种各样数据在传输的社交网站之外，也包括物联网连接起来以后，即便养一头牛、养一头猪都有芯片，都会产生的无数数据。

今天的数据已经不是我们用传统的方式，把数字输入计算机处理一下就得到报表，而是在感知化（instrumented）、物联化（interconnected）和智能化（intelligent）的交会下，就好像把调节水量的 3 道闸门同时开

启一样，遍及各处的数据量，从原本的潺潺细流汇流成磅礴大川，再倾泄灌入一片无边无际的数据海洋。

而企业在这片数据汪洋中安全地航向目的地，还是被巨浪吞没，就是这本书要谈的主要议题，在这个命题之下，这本书里所谈的大数据，并不只是一门新技术，更是以大量的数据为基础，进行业务分析、预估与洞察的创新能力。

在这个变动的环境中、数据爆炸的时代下，全世界经济正在重组。未来，有70%的增长来自新兴市场，会有30亿人口成为新的中产阶级，带动全球供应链发生新的变化，这个重组过程对企业来说，既是机会也是挑战。

如果能把大量的数据，用科学化的方式做到更优化的预估，那么在面对复杂环境所带来的诸多挑战下，不管是企业或政府就有可能运用这些经过提炼的智慧，创造新的增长机遇以及全新的价值。

诚如书中所说，这已经不是一个简单的数据增长问题，而是一场量变形成质变的变革，衷心期盼在巨变之后的新世界里，我们都会是搭上发现者号的一员，持续往创新的方向航行。

钱大群

IBM 大中华区总裁

推荐序二

一个由数据组成的人与世界

“过去两年，人们制造的数据就占了当今全球数据总量的 90%。”

针对这句话，有 3 个观点值得被“特别提出”。

1. 数据增生的速度

大数据以这样的速度逼近到我们面前。如果我们把两年以前的文明累积基数以 200 年为限，意味着我们只用了近 1/100 的时间，就创造了 9 成以上的数据。速度快到我们甚至来不及找出一个确定的中文名词来翻译（叫做：巨量数据、大数据、大资料，还是大数据），我们就急于需要这些新的理解、新的应用，面对新的可能，还要面对新的危险。

2. 个人数据的隐私

从个人的角度来看，我们从未被如此清楚地记录过。我们的基本行为其实没有剧烈的改变，还是要吃穿，还是要交通，还是需要有住的地

方，也还是需要娱乐和教育。即便我们有时候用冷漠来面对，政治依然是我们关心的事，即便我们有越来越长的寿命，用医疗来抵抗死亡依然是我们每天的企图。

但我们从来没有被如此详实地被记录着。我们在什么特定时间去了哪家餐厅、在某一个商场待了 3 个小时，看了一场 3D 动作片，平均消费额超过 1200 元，因为某部新的智能手机上市带动本月多消费了 6000 多元。哪个地点上车，哪一站下车。某市 5 月的 GDP 因为天气比往年多雨少了 14%。购物推荐的算法除了可以推荐商品，也可以应用在类似基因的族群，把你可能会罹患的疾病推荐给你。这些数据，正以新的速度，揭示各种新的理解。

公共领域和私有领域的界限模糊更不只发生在 Facebook 等社交网络，这些数据的归属、谁能用什么样的方式记录什么、如何应用或交易这些数据，正在加剧冲击我们以往所熟悉的隐私，提升或者破坏我们的生活。

3. 数据所带来的权力

我们还缺乏一个允许或不允许、哪些政府或企业、在什么情况下、可以或不可以记录和使用、哪些数据到什么程度的成熟定义与机制。

从另外一个方面来看，拥有数据的组织，会比没有数据的组织拥有更大的权力。拥有数据规模较大的组织，会比拥有数据规模较小的组织拥有更大的权力。拥有较完整的数据的组织，会比拥有较不完整数据的

组织拥有更大的权力。拥有更实时数据的组织，会比拥有更不实时数据的组织拥有更大的权力。拥有消费者信任、主动提供数据的组织，会比缺乏消费者信任、背后搜集数据的组织，拥有更大的权力。

数据被大量地释放出来，权力便被大量地释放出来。知识被大量地创造出来，权力便被大量地创造出来。

缺乏这些数据的组织、缺乏能力将数据转换为知识的组织，就缺乏和这个世界互动的基础，就不能成为这个新世界的一部分。

信息焦虑只是个人遭受信息时代冲击产生的课题，那只是“前菜”。

欢迎来到弥漫“数据焦虑”的大数据时代。

戴季全

《TechOrange 科技报橘》创办人暨发行人

《WIRED 国际中文版》创刊总编辑

目录

Part 3　技术与前瞻



导读

人类生活的下一块拼图

早在几年前，大数据的相关话题就已在科技界发酵，当时大家着重的是技术层面，希望开发更先进的软硬件，更有效地储存、利用这些应对网络时代而不断产生的数据。但是，大数据之所以重要，绝不是更先进的数据资料采集而已，因此每当讨论这个议题时，我总是一再地强调，我们要探讨的主题叫做“大数据分析”（Big Data Analytics）。

诚如本书所提到的，大数据真正的价值在于趋势背后所包含的“分析学”（Analytics），也就是以系统化的方式，把从不同渠道获得的大量数据，转变为经过组织的信息、甚至知识；这就像是一个人同时拥有听觉、视觉、味觉和触觉一样，将这些“感觉”综合起来，才能感知到我们身处的每一个瞬间，然后应对各种状况，做出适当的反应。

尤其是在这一个越来越“平”、越变越“小”的世界里，人与人、人与环境之间的连结越来越紧密，交互影响力也越来越大，人们开始意识到在这个世界上生活，没有人可以置身事外，无论好事或是坏事，就比如说这几年的金融危机、能源短缺、环境污染、食品安全等风险，都已经不再是区域性问题，而是实实在在地影响着各个国家、各个阶层的



每一个人。

同时，通过手机、网络、传感器（sensor）等数字化系统，让人类历史上第一次，几乎任何东西都可以数字化地被测量，各种创新的感应科技大量被嵌入汽车、家电、公路、水利、电力等设施当中，加上网络的高速发展，使得越来越多的人、物品、环境可以被建构成一个互联互通的系统。

从数据里挖出新需求

发展大数据分析的意义是将这些大量增生的新数据，通过分析工具形成新的洞察，进而提早在变化莫测的世界里预知风险，以追求更美好安全的生活，所以它不只是新科技，而是一种新生活形态的前导，而这种新的生活形态，对于人类文明的发展来说意义深远。

如果我们沿时间次序去回顾人类文明的发展进程，目前为止已发生了 3 种类型的革命，背后的原因大都是为了提高生产力。第 1 种类型的分子是“需要花的时间”，分母是“地理上相隔的距离”，人类希望花在交通时间上的数值越来越小，也就是不管距离有多远，花的时间要越来越短，所以我们发明了新的道路系统和运输工具，从马车、火车一直到飞机。

第 2 种类型的分子是“生产量”，分母则是“一定的人口数”。人类希望这个数值可以越来越大，因此本来我们一天只有 8 小时的白天时间用来工作、生产，发明了电灯以后，现在我们一天的生产时间可能增加



到14小时以上。又或者是电冰箱和洗衣机的发明，是因为过去的社会结构，家庭需要女性来照顾，因此过去的女性就被绑在家里，直到有了家电的辅助，女性才有余力投入生产，因此“电气化”在近代史上也是一个很大的改变。

当整个生产的时间缩短、数量变大，我们还希望它质量稳定，因此人类又发明了计算机，来帮助人类做一些重复的事情，例如抄写、填表，也因此计算机最早的应用都是我们现在每天要用的文档和电子表格，之后逐渐发展成可以帮人类处理很多繁琐又精细的重复性工作，像是设备、机床等仪器，借由“设备化”把人为的不稳定性控制到最低。

从利用交通工具缩短时空距离、利用电气产品增加生产量，一直到利用仪器设备优化生产质量，下一阶段人类文明的发展又会朝哪里前进？其实，不管是2000年前、800年前、100年前，人类的基本需求都没有改变，我们追求更美好的生活，所以想办法提升了物质上的需求，但不要忘记，还有另一部分也是亘古不变的，那就是人们心灵上的需求。

从历史上看，今天很多的改变都是为了标准化，以提高生产力，但上帝创造每一个人的天性都不一样，所以不管iPhone卖得有多好，就是会有人不想和大家一样使用它，这种独特性代表着在每一个人的心里面，都有一个需要去填补的空虚，因此除了标准化之外，越来越多企业注意到每一个人内在个性化的不同需要。

举个最简单的例子来说，一间会议室里原本只有两个人，当另外两个



人加入讨论，温度开始升高，这时候空调是不是应该开始调整？调整到几度对人体才会最舒适？如果中途有人离开，或是有人的情绪越来越激动，那么空调系统又该如何应对？满足这些个性化的不同需要，在科技、商业以及社会中都是正在进行中的实验过程（ongoing experimentation），也是人类文明发展中另一个进程的开始。

经历了铜器时代、铁器时代、原子时代和计算机时代，现在我们正在进入爆发的“数据时代”，如果缩小了时间轴（time frame）来看，每一次的大变化都是文明发展中的一次革命，但如果拉长来看，这些革命一样还是在解决人类提升物质和心灵的需求，大数据分析的出现也是如此。

如何利用大量的数据分析、洞察，让人们的基本生活需要以及心灵需要，和这个世界的管理结合起来，是我们希望大数据能够做到的应用和服务，而技术的演进则是帮助这件事情发生的推动力。

拼出未来的样貌

所以，在这本书里我们并没有提供很深入的技术，而是希望以科普的方式让更多人了解，如何利用大数据分析帮助你的企业、组织，甚至生活。尤其是当你可以用实时（real time）或近乎实时的速度，整合来自不同渠道的庞杂数据，并运用强大的计算能力分析和挖掘，那么趋势（Trend）就不再只是像素，而是一块有意义的拼图。

大数据分析可以帮助我们拼凑，甚至预测趋势的发生轨迹，表面上



看来是帮助产业找到新的方向，但是这本书所要传递更深层的意义是，希望这项新技术可以形成一股由数据分析带动的正向循环，让更多人具备相关的知识跟技能，成为这个世界更良善的帮助，引领着你我的生活走向一个更美好的新境地。

李实恭

台达电子技术长

台湾清华大学服务科学研究所兼任教授



Part 1

什么是大数据







第1章 大数据新世界



早上6点半，手机设定的闹铃声响起。

Dave Cheng猛然地张开眼睛，急忙从床上跳了下来。他用掌心搓了搓脸，希望自己快速清醒，然后打开电视新闻台，一边到浴室简单梳洗，一边听听有什么新鲜事发生。

今天应该算是个特别的日子，特别的忙碌的日子！昨晚临睡前他才利用家里的台式计算机，连到公司内部网站上，再次确认今天上午他得参加两场会议，一场部门会议，一场国际电话会议；下午他要拜访一家客户；然后傍晚时飞往上海，陪老板参加科技商展。

走出家门，右转100米，到了公交车站，Dave Cheng检查重要的随身物品，笔记本电脑，带了！护照，带了！20寸的登机箱，右手正拖着！万事齐备。

看了看手表，时针指向7点整，站牌前等公交车的人开始多了起来。3分钟过后，他有些小小的不耐，拿出手机点进"等公交车"APP（手机应用程序），APP上显示公交车再过2分钟就会到站，他这才安心。

场景转到3分钟后的公交车上，乘客有6成以上都是高中生，几乎每一个人都是低头族，时不时听到传送LINE消息时发出的叮叮声。Dave Cheng也打开Facebook，回了几个朋友的搞笑帖子，按了几个赞。

20分钟后公交车到站，他先到附近的便利商店，买了一份三明治加无糖豆浆的8元超值早餐，然后打电话给女朋友说早安。8点钟走进公司、打开计算机，趁着老板还没来，Dave Cheng一边吃着早餐，一边浏览着在线电子报。半小时过去，他读了两篇关于欧债的报导，一篇明星八卦，还顺便查看了一下星期五晚上有什么新电影可看。



然后，他点进公司的专属邮箱，确认昨晚外国客户已经收到了报价单，再看看有没有什么紧急邮件要处理，通常是没有，今天也不例外！8 点半过后，同事们陆续到了，Dave Cheng 打开 MSN 和 Skype，敲了消息灵通的 Tom，问他对于公司最近进行的人事变动有没有什么小道消息，在闲扯中等着 9 点钟的部门例行会议。

截至目前，你是不是觉得有些平凡无趣？的确，Dave Cheng 的生活和一般上班族没什么两样，但你没注意到的是，光是从起床到公司上班的这两个小时内，他平凡的个人行为，已经产生了数亿位的数据，而且随着时间过去，还在不断增加中。

不断增生的数据巨浪

使用手机 APP 程序、上网到 Facebook 按赞、或者是打电话，都有无数的数据资料跟着产生，即使是到便利商店买东西，店内的 POS（Point of Sale，销售点终端）系统也正记录着每一笔消费信息，更别提他上网漫游时所点击的网页、MSN 回复的消息，这些轨迹通通都是数据资料。

在这个数字化的世界里，一天 24 个小时，一年 365 天，人们无时无刻都在生产着各种数据微粒。只要是使用手机、计算机、信用卡……都会产生并传送出无数关于我们的数据，就算是用移动电话发送一个笑脸符号给朋友，在你还没来得及把手机放回口袋中的一两秒间，这个小小的动作就已经穿过光纤网络，发射到卫星上，抵达远在数千公里外的某个服务器里，变成数据记录下来。



只要活在世上，每一天，我们“生产”的数字档案就会越来越多。如果“位”肉眼可见，那么我们每一个人都像是采蜂人一样，身上附裹着一层又一层、厚厚的“位”数据。

再继续观察下午3点，正坐在客户公司会议室里的Dave Cheng，从两个小时前银行保安监视摄像头捕捉到的影像，以及ATM（自动柜员机）的记录上，可以知道他提取了5万元；从一个小时前信用卡的签单记录上，可以知道他喝了两杯咖啡，而移动电话公司利用他的通话记录定位，发现他的活动范围几乎都在东区，短短两个小时，又是好几亿位数据的产生。

大数据新世界

我们所面对的，不只是一群随时产生数据的个人，更是一个不断被数据淹没的大数据新世界。

每一秒，一家大型医院会产生12万笔生理健康数据；每一分钟，YouTube网站上传影片的总长有72小时；每一天，一家银行要处理500万笔信用卡交易、一个Twitter网站上有2.3亿条tweet。

如果再加上全世界同一时间约有超过5亿部智能手机、10亿台计算机和数万亿个传感器同时运作，所产生的各种文字、声音和图片数据，每一天制造出来的数据量估计高达25亿GB（吉字节），等于要用4000万台64GB的iPad才能装载。而且，光是过去两年，人们制造的数据就占了当今全球数据总量的90%。

依照这种每年约50%的增长速度计算，科技研究公司IDC估测，



到了2020年全球数据总量将增长44倍，达到35.2ZB（泽字节，相当于1万亿GB）。如果把这些数据全都装在64GB iPad里，这些iPad迭起来的高度足足可堆出超过13万座玉山（位于中国台湾省中部，海拔3997米，是台湾地区的最高峰）。

ZB到底有多大？有人形容ZB等级的数据量就像全世界海滩上的沙子数目一样多，更惊人的是，它还在不断蔓延！如果以每一分钟在网络世界流动的信息量来看，当你打上关键词，按下“Google 搜索”的这一刻，其实你只是200万人中的1人；当你写好电子邮件，按下“传送”的那一刹那，这封电子邮件也只是2亿封其中的1封。

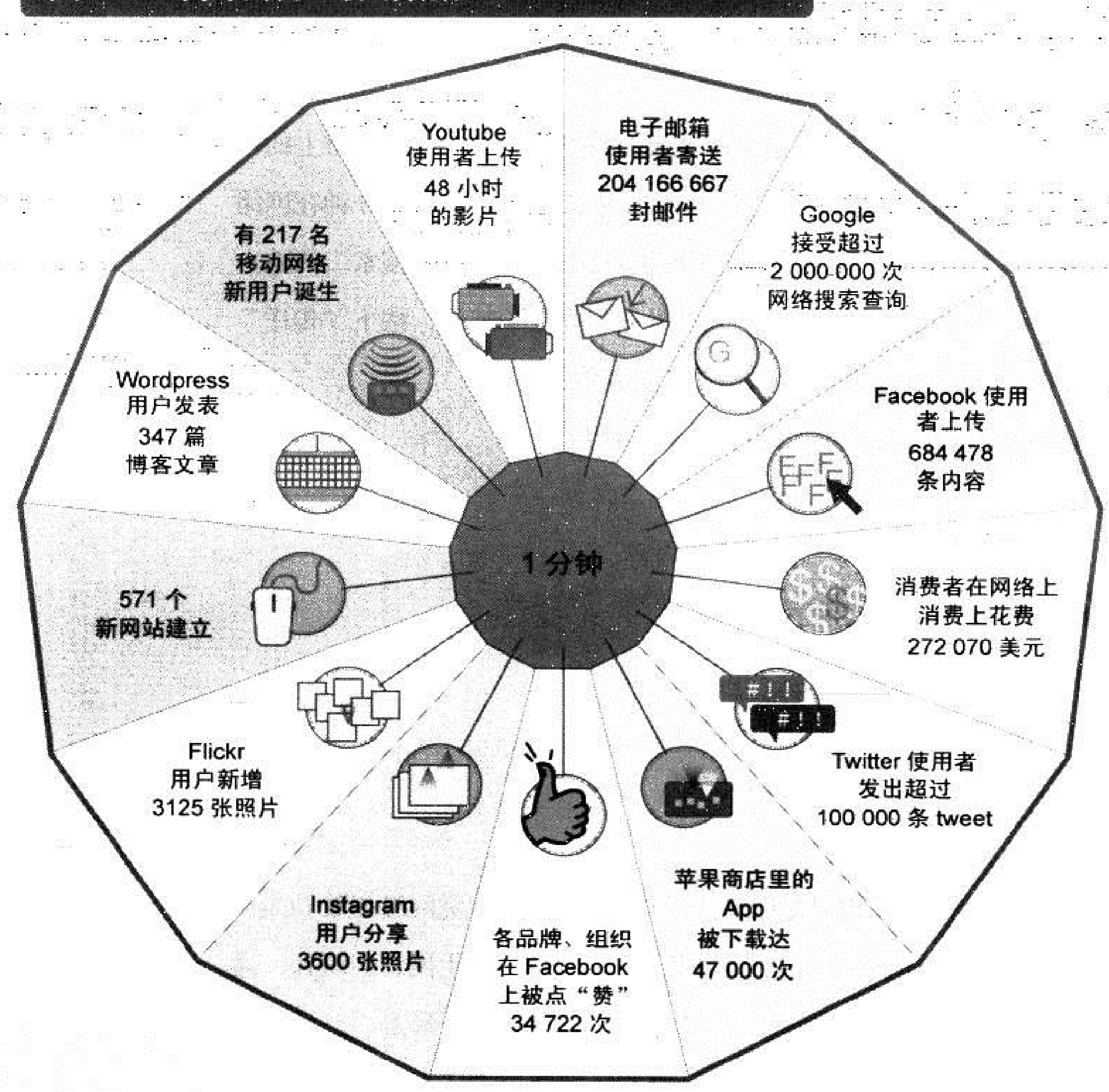
其他更惊人的一分钟网络数据资料包括：

- Facebook上产生超过68万条内容；
- 超过27万美元的网络购物交易；
- 苹果商店里的APP被下载达47 000次；
- Flickr用户分享了3125张照片；
- 有217名移动网络新用户诞生。

请记住，上述的数据只是现况，而未来呢？思科视觉网络指数（Cisco Visual Networking Index）发现，以前人们只用计算机上网，但现在，通过计算机、手机、平板电脑等多种设备随时上网的生活形态，已逐渐成为文明世界的常态。2011年，全球网络联机设备为103亿个，以地球上的70亿人口计算，每个人分配到1.4个，2016年前网络联机设备总数将

增长至 189 亿个，等于每人约有 2.5 个。

图 1-1 持续增长的大数据



在中国台湾省，每个人拥有一台台式计算机、一台笔记本电脑、一



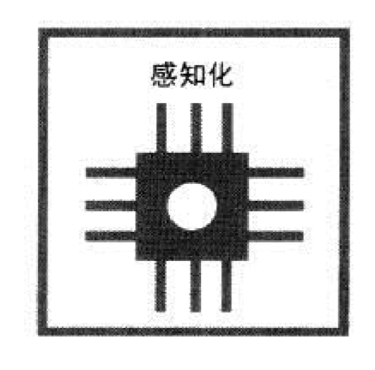
部智能手机，是再普通不过的基本配备，现在可能还要加上一台平板电脑。而人们“随时在线”也让全球网络流量正以一年增长 1ZB 的速度增加当中，2016 年前将达到 1.3ZB，平均流量将达到每秒 245TB（太字节），相当于有 200 万人同时观看高清影片。

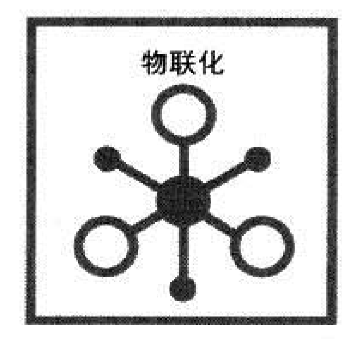
为什么数据量暴增的速度会这么快？因为这个世界已经变了，变得感知化（instrumented）、物联化（interconnected）和智能化（intelligent）。

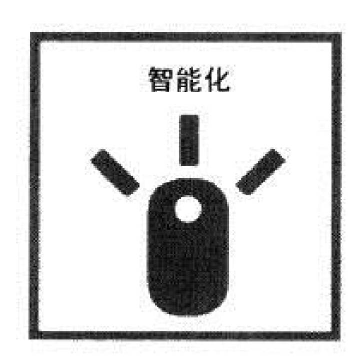
简单来说，所有的物体，包括风、流水、空气中的湿度，都能被感测，这就是感知化；感测过程中产生大量的数据，需要输送到后台进行处理，这就是物联化；而获取数据只是一个手段，最终目的是要从庞杂巨量的数据资料中，分析出有用的信息，帮助人们做出决策，这就是智能化。

而这 3 个“i”，也分别反映了数据源、传送方式和使用方法的改变。

图 1-2　大数据的 3i 新世界

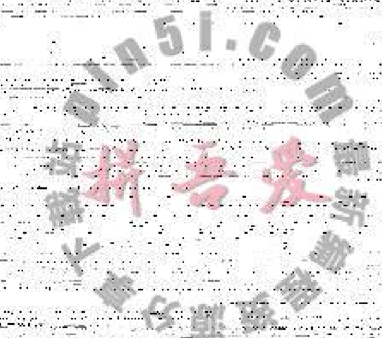






感知化的世界

“感知化”指的是数据源的变化。你或许没有感觉到，大量内建芯片、



传感器、RFID（无线射频芯片）等具有“电子神经”的感知设备产品，其实已经遍布在我们的生活周围。2011 年，全球嵌入的芯片、传感器、RFID 卷标的电子设备，初估超过一万亿个，世界上的每一个人平均约拥有 10 亿个芯片。这些无所不在的电子设备也是增加速度最快的数据源。

电子设备与物品相互连接、成为网络的“感知化”现象，使得人们可以更灵敏地、更全面地感知物理的世界，促成这个改变的最大原因是晶体管（transistor）技术的突破与普及。1947 年晶体管问世，为数字时代揭开序幕，直到 20 世纪 80 年代，上面覆盖着数百万微小晶体管的小小硅芯片都还是新奇的发明。

不过，科技设备发展也的确如同摩尔定律的预测，单一硅芯片的晶体管数目，每隔 18 到 24 个月增加一倍，芯片容量也会增加一倍，且成本等比例下降。循着摩尔定率发展的半导体业，让晶体管的体积越来越小，芯片价格也越来越便宜。

以计算机的大脑——微处理器来看，1994 年掀起笔记本电脑风潮的 Intel-486 芯片，单价约 300 美元，当时搭载了 800KB 容量的晶体管；而 10 年后，Pentium 4 3.0 芯片，搭载 55MB 容量的晶体管，数目是 486 的 68 倍，速度是 486 的 120 倍，现在的售价却比当时的 486 还要便宜。

如今，一个芯片里面已可包含数亿个晶体管，平均每个芯片的成本只要十万分之一美元。售价只有几美元的音乐生日贺卡，其计算效能早已超越了数十年前最快的大型主机，从个人计算机、移动电话、电视游戏机，到汽车（内含 GPS 导航系统）、宠物项圈（内含身份识别功能），全都有这些便宜芯片的身影。



越来越多芯片被“植入”我们的生活之中，记录它们接收到的每一个指令。其中，也有极大量的芯片被制成传感器与电子卷标，再置入到监控摄像头、大楼温湿度传感器、医院检查仪器、风力发电机和大卖场中的无线射频识别系统（RFID）中，各种各样、总数上万亿的电子设备，不受时间限制地感测着人们工作、购物和休闲的各项动态。

例如这两年来，全世界的电力公司为了节省能源，戮力研发的智能型电表，就运用了大量的传感器，一天24小时不停歇地测量、提取和传输终端用户的用电量信息。

对消费者来说，这些电表实时反映出家里的耗电量与电费信息，也可依据现阶段用电量估算未来24小时的电费会增加多少，有助于用户控制和调整用电的习惯；对企业而言，电表则变成远程的监控传感器，帮助它随时掌握电网供电的状态，万一耗电量瞬间飙高、可能超过电网负荷时，电力公司就能提早采取应变措施，大大减低了无预警停电的概率。

电表、水表、煤气表，这类智能型仪表已大量被嵌入到全世界各类的器具中，不仅装设数量惊人，而且很快就会存在于每一个人的家里，24小时记录着每一个用户的能源消耗量，然后捕捉、测量和传递数据。

为了不间断地计算流量，一个智能型仪表需要至少每秒检查簧片开关状态20次，至少每15秒钟就要建立一个无线数据封包，并将这些数据传输到无线发射器进行传送，这也是目前最难以计数且大量增加的新数据源。



物联化的世界

“物联化”指的是数据传送方式的变化。根据联合国公布的统计数字，目前全球网络的使用人口已经突破20亿大关，使用手机的人口更已突破50亿，而计算机、手机等3C产品都可以和前面提到的上万亿个，可能来自汽车、电器、道路、自来水管，甚至是食物包装盒的传感器，彼此链接并交换信息。

当机器与人类社会系统全面互连互通时，所创造出来的各类数据量非常庞大，也难怪有分析师估计，“M2M”（machine-to-machine，机器对机器）每年可带动超过两位数的数据量增长。

M2M的运用主要是通过移动通信（手机）为核心，对设备进行有效控制。从狭义的定义来看，M2M是代表机器和机器之间的自动通信，而且不只是简单的传输数据而已。换言之，即使人们没有发出信号，机器也会根据既定程序主动进行通信，甚至根据所得到的数据做出筛选后再传输。

以这些数据的流动量来看，2010年底时还只有3%的网络流量来自电视、平板电脑、智能手机以及机器对机器（M2M）模块等非PC设备，但预计到了2015年时，非PC的网络流量将增长到13%，其年复合增长率分别为电视的101%、平板电脑的216%、智能手机的144%、M2M模块的258%。以增长潜力最大的M2M来说，届时全球会有150亿台机器，可以不通过人工的介入直接互联。



目前，M2M大多应用在远程监视、控制、以及数据追踪和供应链管理上。例如，为了应对层出不穷的食品安全问题，山东省商业集团正在导入一套猪肉生产追溯管理系统，这套系统广泛使用了监测设备与运送猪肉的多元数据，从生产源头农户、肉品加工到零售店的整个商品流通过程全面互连。

这套系统链接了条形码、温度和湿度传感器以及全球卫星定位系统（GPS）等不同技术的多元数据，可以对生产、流通和零售等各阶段的猪肉产品状态统一管理，如果消费者因食用猪肉而健康受损，可以立即确定销售猪肉的店铺，也就可以尽快实施包括下架、回收等应对措失。

首先，在生产阶段，屠宰场内装设RFID设备，收集猪从进场到宰杀，以及运送前各个流程处理的数据，保障肉品处理流程的效率，也避免因为某环节的疏失（如某批肉品未及时装运）而影响肉品质量。

其次，在运送阶段，运送货车采用温湿度传感器、GPS 与地理信息系统（GIS），每隔一段时间传送货品温湿度和所在位置回中心主系统。要是货柜内温湿度不符合标准，系统便会自动提醒负责人员采取相对应的行动，及时排除问题肉品进入销售链的可能。

最后，在零售阶段，超市结账人员通过产品上的条形码和柜台的收银系统——扫描记录每件生鲜猪肉商品销售的时间和地点，如果真的有消费者食用猪肉后生病而通报，主管机关就可循着线索找到贩卖猪肉的店铺，并且尽快召回同批出厂的问题产品。除了猪肉之外，这套系统也将运用在海鲜和鸡肉等生鲜的生产管理上。

随着装载微型芯片的电子设备互连日益紧密，M2M的定义已经扩展

到人对机器（man-to-machine）、机器对人（machine-to-man）之间的通信，数字世界和实体世界的界线逐渐模糊，越来越多和计算机看似毫不相干的东西，现在都能够具备计算机计算能力并且传送数据。

举例来说，体积轻薄的微型计算机被广泛应用在许多领域上，例如手表里插入信用卡芯片，购物时可以直接刷“表”，而不是刷卡；或者有心血管疾病病史的患者戴上可监测心跳血压的手表，一天24小时随时传送数据资料，一旦指数发生异常，系统可以立刻发送消息给附近的医疗院所，让患者及早获得必要的协助。这些可以互相“沟通”的机器正发展出新的应用层面，也正在产生更大量的数据。

智能化的世界

“智能化”则是指数据使用方式的变化。当上述感知化、物联化的网络被注入强大的分析、计算能力后，各种设备、机器具有比以往更高的人工智能，也因而改变了数据使用与处理的方法。

在计算技术的创新突破下，散布在四面八方的终端电子产品和传感器等各种设备，和后端的计算机链接之后，数据被大量地、系统性地解构、处理，再加上新型计算架构的兴起，如云端技术，或是把传统计算机集群起来形成的平行计算架构的Hadoop（参考第13章），就能够整合和分析跨越不同地理区域、产业和领域的大量数据，进行复杂的分析、统整、演算和预测。

以2011年在美国知名的益智抢答竞赛《危险境地！》（Jeopardy）中，



打败人脑的超级计算机“华生”（Watson）为例，在经过 3 天激战、一次宕机之后，它最后击败了该节目史上两位最强的高手詹宁斯（Ken Jennings）及拉特（Brad Rutter），赢得 100 万美元奖金，也改写了超级计算机的历史。

华生是以 2 800 个处理器核心、16 万亿字节的工作内存运转，每秒计算能力高达 80 万亿次。得知问题之后，它得先针对句子中的名字、数据、地理位置或其他条件，运用 600 万条逻辑指令层层分析才能找到正确答案，而找到答案之后，还要快速控制金属手指按铃抢答。

而且，即使华生内建了 2 亿页、4 万亿字节的百科知识库，但它不仅要熟知重要的历史人物、文学、科学、艺术、娱乐及游戏策略等知识，还要了解笑话、双关语、讽刺语及谜语等隐喻，才有可能答对这些复杂甚至藏有陷阱的题目。

我们以比赛中的题目之一为例，“以色列的摩西 · 达扬（Mosche Dayan）是以什么装饰让全世界都认识的？”以计算机的计算方式来说，华生要先确认“达扬是一个地方吗？还是一个人名？或是一处《圣经》提过的圣迹？”之后它要从以色列军队、名人语录、甚至穿着风格等来判断这是一个人名，再从数百种可能的答案中挑出一项正确答案。

这位以色列前国防部长达扬将军的左眼眼罩是他的著名象征，华生必须在 3 秒钟以内回答出正确答案才有机会获胜，而华生做到了。你可能不知道，这个益智节目中所抽样的两万个问题中，有高达 2500 种题目类型，而要回答其中一个简单的问题，一般计算机则要花两小时，原因就在于听得懂人类语言、能够和人类谈话对计算机来说难如登天。

要计算机了解语言变化远比关键词搜索困难太多了，因为人类在日常沟通时就常常语意模糊、不精确。以一道华生答错的题目为例，题目是“有一本小说的书名及 1957 年一部电影的片名灵感都来自于这个东西，且它横跨湄公河（Mae Khlung）”。答案是《桂河大桥》，但其实华生没答对是很合理的，因为其实桂河大桥并不如题目所说的横跨湄公河，而是建造在桂河上，而桂河则是湄公河的支流。

我们不能说计算机已经打败了人脑，因为华生还无法在词汇有谬误时精准地排错，在比赛过程中的按铃速度也常常比人类对手慢上一步，但它的确证明了机器不仅能收集、储存庞大的数据，还已经能像人类一样思考。事实上，我们日常生活里拥有的终端设备，也已经有了不同程度的“智能”，判断哪些程序或功能的使用状况，或是或该用什么方式传送与处理数据。

让数据不只是数据

这就像是在智能手机安装了各种各样的 APP 应用程序，有些 APP 需要传输和处理数据，例如天气预报 APP 需要链接气象局的数据系统，或是运用卫星定位系统呈现用户所在地的卫星云图。相比之下，手机里内建的单机小游戏或是记事本等程序，就不必传输或处理这么大量的数据。

智能手机之所以“智慧”，是因为它有判别能力。在处理过程中，不是每笔数据都值得分析，也不是每一次分析数据都要动用到全部的计算力，因此在数据送入分析之前，数据源头的设备和数据汇集的节点

（Node），如果可以有智慧地“清洗”一些不需要或不合格的数据，再送往后端平台中进行处理，分析和演算的效能和结果，参考价值将会更高。

以银行服务为例，同一个客户在某银行的网络平台购买基金，也在同一家银行柜台存款，对于银行的核心计算机来说，网银和柜台系统的设备就像两条并行的神经线一样，各自传导着同一个客户不同的交易信息。如果在这两条神经线连结时的节点，就可以去除数据的杂质且分门别类，主核心计算机的分析效能会更高。例如，客户英文姓名统一采前姓后名的格式、客户 ID 都以身份证号而非出生年月日为准、中国台湾地区的邮政编码统一采 5 码而非 3 码的格式等，再连到大脑（主核心计算机），大脑不用浪费时间和效能进行重复比对，就知道这是同一个人的数据。如此一来，数据分析出来的结果也会更即时、更准确。

如果机器无法智能化，即使有了“感知化”与“物联化”的设备，被产生、收集和传送的数据也只能被储存下来，而无法用来分析和辅助决策，那么，数据就永远是数据。但也是因为机器被赋予了更高的智慧，产生了更多的思考结果，更多的数据也随之而生。

时代变革的起点

在感知化、物联化和智能化的交会下，就好像把调节水量的 3 道闸门同时开启一样，遍及各处的数据量，从原本的潺潺细流汇流成磅礴大川，再倾泄灌入一片无边无际的数据海。

在如此巨量的数据冲击下，这已经不是一个简单的数据增加问题，

而是一场量变形成质变，足以匹敌20世纪科技革命的巨大变革。

2012年2月，物理学家迈尔斯（Mark P. Mills）和美国西北大学应用科学院长欧提诺（Julio M. Ottino）在《华尔街日报》撰文，认为100年前出现了电气化、电话、汽车、不锈钢和无线电放大器等机器，改变了人们的生活方式。时隔100年，我们再度站在时代变革的起点。而再次改变世界的推力之一，就是伴随着无数机器而来的大数据分析。

哈佛大学量化社会科学学院（Institute for Quantitative Social Science）院长盖瑞金（Gary King）认为，庞大的新数据来源所带来的转变，将在学术界、企业界和政界中迅速蔓延开来，没有哪一个领域会不受到影响。而巨量数据资料的处理能力将使以往无法想象的服务和业务成为可能，进一步改变人们的生活方式，甚至引领新一波的经济繁荣。

2012年的伦敦奥运会才落幕不久，令大家惊艳的不仅是精采的赛事，还有历史悠久的英伦之都，如今已运用大数据蜕变成一个智慧城市。

拥有149年历史的伦敦地铁，11条路线全长超过400公里，沿线建有270座车站，每年运送10亿人次。为了让地铁干线正常运行，伦敦地铁里的每辆火车都有GPS，站台上的乘客随时可以在显示牌上了解下一趟车的抵达时间；站台上布满无数的传感器，将等候的乘客人数提供给控制中心，让调度人员可以灵活控制车次和出车时间间隔。

同时，为了快速得到最新的流动信息（例如进出站人数、等候人数），伦敦地铁站内也铺设了无线局域网络（Wi-Fi），而现在，任何人都可以在地铁站里通过免费的Wi-Fi，用手机查看地铁实时动态地图，以及接收各种地理位置的便利信息；博物馆、艺术中心、歌剧院、或是酒吧也都有



相应的地理信息 APP，各种文化、艺术、科学等资源，也全都可以通过无线网络获得丰富而详尽的免费数据。

从地铁站走到街上。伦敦是欧洲第一个对汽车进入市中心要额外课税的城市。由于交通拥塞，人们可以在停车高峰时段，用手机上网实时查询停车位的空置情况，并且下单预订停车位，因为只要随意停车，伦敦交警会用掌上电脑在你车前的条形码上扫描一下，包括这辆车的车速、停车记录、未缴费记录等各种信息全都无所遁形，几秒后罚单就从警察手中的小型计算机里直接打印出来。

另一方面，为了应对奥运期间涌入的百万游客，伦敦街头设置了超过 100 个配置液晶屏幕的智能型垃圾桶，与 Wi-Fi 相连，不仅可以指示民众如何分类处理垃圾，还可以显示天气、气温、时间、股市行情等，里面内藏的微型摄像头也可以防止街头犯罪和恐怖攻击。

伦敦以大量数据和科技设备建构了一个数字化的智慧城，但伴随而来的则是更大量的数据，为此伦敦市政府建立一个城市网络数据中心。每一个公务员都要把公共数据丢进这个包括交通、安全、经济发展和旅游业的开放式数据库里。民众可以从这里取得交通拥堵数据的实时更新、地铁业务指南，或是自行车出租计划的取车地点分布，也可以自行经过切割和分流，将这些数据放到自己的个人计算机和其他电子装置里，进行商业开发。

例如有一个名为 WhereDoesMyMoneyGo.org（我的金钱去向何方）的网站，就专门追踪民众税金的流向，而这也使得伦敦公共建设的成本和中标价格非常透明。伦敦市政府认为，把这些数据移交给能够把事情



做好的人，比城市直接提供那些服务的成本更低，后来甚至推出电子商务化的数据商店（Data Store），向开发者提供多种 API（Application Interface，应用程序编程接口），激励创新开发。

新世界的新竞争力

成功运用大数据分析打造新风貌的伦敦，已成为全英国未来 10 年内发展的重要依据。英国智库政策交易所（British think tank Policy Exchange）在 2012 年 6 月发布报告认为，大数据分析可为英国政府提高效率及削减浪费，一年可能省下 160 亿～330 亿英镑。

美国奥巴马政府也将大数据分析视为下一步的国家发展战略，白宫在 2012 年 3 月宣布投资 2 亿美元启动"大数据研究和发展计划"，包括大数据分析以及大数据在医疗、天气和国防等领域的应用。白宫甚至将数据资料定义为"未来的新石油"；换言之，一个国家拥有数据资料的规模和解释运用的能力，已成为一个国家的核心资产和国力指标。

数据分析运用的重要性对国力如此，对企业竞争力更是如此。在零售业，美国的沃尔玛公司很早就开始利用事务数据库来赢得竞争优势。1969 年沃尔玛开始使用计算机来追踪存货，1983 年所有门市开始采用条形码扫描系统，每一样商品的"身份"都可以存进计算机数据库，1987 年完成内部卫星系统，汇整全美各分店的实时数据，借此分析顾客的购买行为。

来年，数据资料就帮助沃尔玛成就了一则零售业的经典传奇。当时，



管理人员分析销售数字时发现了一个令人难以理解的现象：啤酒和尿布这两件毫无关联性的商品，销售数字确有着难以理解的高度正相关，尤其是在年轻爸爸的购物车里。后来发现，这是因为年轻夫妻的分工形态，通常是由妈妈在家照顾小孩，而爸爸外出购物，而买小孩尿布时，爸爸们通常也会带个自己想喝的啤酒回家。

沃尔玛发现了这个独特的现象，开始在卖场尝试将啤酒与尿布摆放在相同的区域，让年轻爸爸可以同时找到这两件商品，并且很快地完成购物。调整之后，结果尿布和啤酒的销售量双双增加 3 成。2005 年卡翠娜飓风来袭之前，沃尔玛也用相同的逻辑，从手电筒和电池的销售数据中分析出馅饼将会热销（因为飓风来袭时导致停电，所以人们特别喜欢方便食用的馅饼）而把手电筒和电池货架移到冷冻柜旁，业绩果然也如预期增长。

西班牙品牌 ZARA，则是运用数据分析引领快速时尚（Fast Fashion）风潮的崛起。每天，ZARA 平均卖出 110 万件衣服，通过全球信息网络，每一件销售出去的商品都有自己的销售身份证（包含售价、部门、时段、客户层），这些数据经过自动化程序分析出顾客的行为模式和消费喜好，做为产品的生产决策，让 ZARA 最短 3 天就可以推出一件新品，一年可推出 12 000 款时装。

在职业运动业，全世界获得主要洲际赛事冠军最多的球队 AC 米兰，也利用每一场赛事的影音文件分析球员数据，进行球员的运动损伤预防和治疗管理，精准度高达 70%，而在一个完整赛季中，因为球员损伤而无法比赛的天数减少了 2/3。后来，美国职棒的波士顿红袜队、旧金山巨



人队和密耳瓦基酿酒人队开始跟进，甚至依此模式发展了一套3D影像特训法，把对手和自家选手的图像文件变成3D显像，可以从任何方向观看、向前转和倒转，并加以解析，打者可以反复比对同一投手的不同球路，或是利用3D眼镜和拟真的动画对手（动作依特定真实对手的统计数字量身打造）对战，以增进球员的战力。

而在公共领域中，数据分析的预测能力也正在开发当中。4 年前，Google 和美国疾病控制及预防中心合作，以关键词搜索次数发展了 Google Flu Trends，协助"追踪"流感传播趋势，至少可以提早两个星期掌握流感爆发的关键时刻。哈佛大学最近一篇医学研究报告也发现，通过 Twitter 监控海地的霍乱疫情比以传统方式监控来得更为有效率，如果再配合政府与金融机构的数据更可以发现食物与水资源短缺的早期征兆。

联合国"全球脉动"（UN Global Pulse）研究计划，更进一步利用社交网站、网络论坛和博客帖子进行"情绪分析"，预测失业率。研究中发现，失业率上升的 3 个月前，网络上关于就业问题的抱怨或沮丧发言就会开始增加，而在失业率上升后的 2 个月和 3 个月，则是房屋亏损和缴不起车贷的话题会增加，此时房地产业、汽车业的购买人气也开始受到影响。

人类科学的范式转移

大数据分析不仅正在改变我们运作企业、制定决策、创新商业模式、

管理风险的方式，同时也推动人类科学研究进入一个新范式。

为了纪念发明数据库的著名科学家格雷（Jim Gray），微软出版了《第四范式》（The Fourth Paradigm: Data-Intensive Scientific Discovery）一书，书中这位曾经获得被视为信息业诺贝尔的杜林奖得主认为，科学发展已经走过了“实验、理论、计算”3个范式，渐渐形成以“数据”为重点的第4范式。

他主张，人类科学研究的历史划分为4个阶段：几千年前是实验科学，主要是描述自然现象；过去几百年是理论科学，描述的是物体运动现象的牛顿定律，或是描述电磁现象的马克斯韦尔方程组（Maxwell's Equations）；而过去几十年，转移到了计算科学，就像是之前所说的，以超级计算机仿真复杂的各种现象。

但到了今天，新范式是数据密集型科学，也就是理论、实验和模拟的汇整。未来的科学发展，将取决于不同学科的研究者如何彼此合作，运用密集数据技术，改善处理流程，并通过云计算的分散平行处理技术、可视化方式，来分析、提炼、解读数据。

例如，微软正在研究一款从过滤垃圾邮件而来的数学模型，因为科学家从筛选垃圾邮件的过程中，发现这些垃圾邮件有类似“突变基因”的设计，以便躲过各种新的过滤方法，而这和HIV这类引发出艾滋病的突变型病毒的变化轨迹有相似之处，也许可以从这方面找到HIV病毒的有效疫苗。

通过数学演算，心理学家、经济学家、生物学家和计算机科学家，正以前所未有的方式密切合作，从我们生活点滴中涌出的庞大信息，以



全新的视野开发下一个撼动人类文明的新发明。

我们可以预见，21 世纪最伟大的发现之一，将来自于从庞大的数据资料中找出的新型态；21 世纪最伟大的工程之一，将是仿真人性的数学模型建构。在这个由数字、向量和算法构成的大数据新世界里，整个世界就是创新和发现的人类行为实验室，不管你愿不愿意，置身其中的你我都是参与者，但如果你愿意从现在开始了解大数据带来的冲击和影响，你就不会只是参与者，而会是下一个发现者。





第2章 不只是大而已

随着环境、技术、生活形态的转变，这个世界正在快速累积大量的数据，或许你会纳闷，在商业世界中，从大量数据里挖出新规则，一直都不是什么新鲜的东西，那么大量的数据和我们所说的大数据分析，到底有何不同?

1997 年击败当时国际象棋冠军卡斯帕罗夫（Garry Kasparov）的超级计算机“深蓝”，和上一章我们提到在益智抢答竞赛中打败人脑的“华生”，或许可以帮助我们诠释大量的数据和大数据分析的不同。

在人类努力发展计算能力下所诞生的“深蓝”，是通过将象棋的游戏规则转化为 0 和 1 形式的算法，扫描数据库后将结构化的查询与答案配对，计算出下一步的走棋策略。当时的深蓝把计算机长于记忆与计算的优势发挥到最大，不但能记住成千上万局的高手棋谱（大量的数据），还能通过计算比最优秀的棋手多预测好几步棋，所以占了上风。但“深蓝”只能够按指令行事，并无独立思考的能力。

而计算机“华生”呢？它不但能理解主持人的问题，还能在短短几秒钟内找出答案，赶在两位超级冠军前按钮作答，这里面不仅牵涉到对人类自然语言（而非计算机程序语言）的理解，还能读取百科全书、报告、报纸、书籍等大量的人类知识，并且从中评估证据、建立假说、评比推理，在短短数秒内计算每一个答案的可信度，最后从中选择一个可能性最大的正确答案，难度可是比下棋高出了太多。

虽说“华生”存放了包括维基百科以及各种字典、文学作品、网络文章及数据库的内容（大数据），但重点是如果不能在 3 秒以内，从相当



于 2 亿页面的数据大海中捞出针来，也不可能获胜。这也就是说，以往计算机分析系统处理的是标准化、结构化的数据，但这些数据通常只占到所有数据的很小一部分，大部分的数据广泛存在于社交网站、电子商务、传感器等散见于各处的数据。

所以，如果我们只按照字面上的说法，认为大数据仅在于强调数据量的巨大，其实是过于简化了它所带来的冲击和挑战。大数据分析，并不是指过去的数据量"较小"，而现在的数据数量"超大"而已，庞大的数据量只是其一，和过去相比，大数据分析还包括数据种类的复杂度和传送速度的增加。

不只是大而已

严格来说，"大数据"到目前为止，还没有一个统一的定义，大多数的说法是，"超过典型数据库工具的硬件环境和软件工具所能获取、存储、管理和分析能力者"即被视为大数据。更简单地来说，"大数据"一词指的是无法以传统流程或工具所处理、分析的数据。

为什么以往的数据处理方式无法处理大数据？因为在这些数据中，除了少部分是结构化（structured）数据外，其他绝大多数都属于半结构化（semi-structured）与非结构化（unstructured）数据。

结构化数据，指的是具有明确关连性定义的固定结构数据，也就是经过编码后存放在数据库应用系统内的数据。在以往的数据库应用上，"数据"必须完全以明确的预定格式被存放，通常是以表格的型式呈



现。以股票事务数据库为例，数据库收到的第 1 个字段数据被设定为 MM/DD/YYYY（格式中的日期），第 2 列是一个 12 位数的数字账号，第 3 列则必须是 3 到 5 位字符字段的股票符号。

也就说，数据库中的每一笔数据都要以事先设定的格式，并按指定的顺序出现（可以表格的型式出现），但这只是结构化数据的第一步，接下来还需要进一步让数据的结构更清楚。我们以 A 公司的人力数据库来说明，员工数据前 5 列分别是姓名、电话、职位、薪资、奖金，初步编码的呈现形式为：

Name（姓名）=Input Name（输入姓名）

Phone（电话）=Input Phone（输入电话）

Title（职位）=Input Title（输入职位）

Salary（薪资）=Input Salary（输入金额）

Bonus（奖金）=Input Bonus（输入金额）

在输入数据时，得依以上的设定输入，才能让数据库“认得”这是需要处理的“原始数据”。

不过问题来了，虽然这是每一个员工的共同数据，但是其中的“奖金”栏却只有主管职级才能领到，所以要把这些原始数据结构化，我们必须把程序代码改成：

Employee.Name（姓名）=Input Name（输入姓名）



Employee.Phone（电话）=Input Phone（输入电话）

Employee.Salary（薪资）=Input Salary（输入金额）

Employee.Title（职位）=Input Title（输入职位）

Supervisior.Bonus（奖金）=Input Bonus（输入金额）

如此一来，我们才可以将员工的薪资和主管奖金分开计算，等到须从数据库采集数据、进行薪资结构和人事成本的分析时，判别出各个变量之间的意义与关连性。

半结构化数据则是属于非纯表格型式、也非纯文本型式的数据，例如 XML 或 HTML 格式的网页数据、电子邮件和办公处理文档等。虽然半结构化数据已有程序编码既定的逻辑和格式，但不容易被数据库理解，尤其是内容含有许多不必要的噪音和混杂了不同的格式。

博客就是半结构化最好的例子。为了排版和阅读方便，博客的确有相关字段、分隔符等有逻辑的程序编码，你可以依照“由近到远”的方式排列文章，也可以依照主题将文章分门别类，或是在固定的位置放上照片或影音文件，这是结构化的部分，但在博客中这些数据并不会互相追随，以固定的模式产生关连性。

例如，你可能不是每篇文章都会放上照片，也不是每天都会放上影音文件，而且即使是文字，今天感触良多写了 1500 个字，而明天可能无事发生只写了 20 个字。或者，应该是纯文本的部分，却多了其他形式的描述，例如在文末放上一个笑脸小图案，或是图释；这些都会影响未来分析判断的准确性。



非结构化数据是指没有固定格式、难以以统一的概念或逻辑分析的数据，这类数据包括文件、图像、声音、影片等。以文件为例，就有纯文本档、Word 文件、PDF 文档等不同的格式。

图 2-1　IT 的数据形态

结构化数据 (Structured Data)	数据库数据
半结构化数据 (Semi-structured Data)	电子邮件、博客文章等
非结构化 (Unstructured Data)	文件、图像、声音、影片等

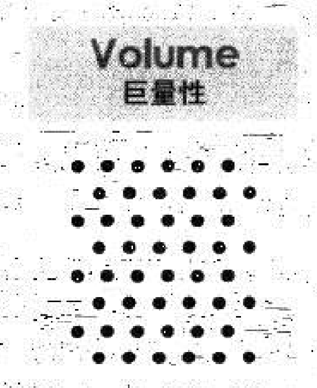
事实上，大数据分析之所以开始被广泛讨论，很重要的原因之一就是，现在无论在哪个领域，所产生的大量数据大多是半结构化或非结构化的，即便有办法储存，提取这些数据之后若要分析运用，实在是很不容易，这也是人类发展信息化的数十年间未曾预想到的情形。

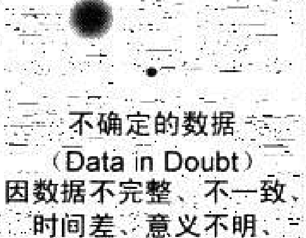
如今，大数据分析的 4 大特性：数据量庞“大”（Volume）、种类繁“杂”（Variety）、变化飞“快”（Velocity）和真伪存“疑”（Veracity），已让政府组织、学术单位和企业面对排山倒海而来的数据巨浪时，开始意识到快速蔓延的大数据分析是已经开始，且永远不会消失的挑战，因此对大数据分析的认识必须更深、更广，才能建构一套有别于以往的技能，补强我们现在数据处理的方式，提升我们对既有知识领域的掌握。



图 2-2 大数据的 4 个 V

Volume 巨量性	Velocity 即时性	Variety 多样性	Veracity 不确定性
存放的数据 （Data at Rest） 数 TB ～ EB 的已存在、待处理的数据	流动中的数据 （Data in Motion） 流数据、反应时间仅有短短几秒～一微秒	种类繁杂的数据 （Data in Many Forms） 结构化、非结构化、纯文字、多媒体数据等	不确定的数据 （Data in Doubt） 因数据不完整、不一致、时间差、意义不明、蓄意欺骗而导致的不确定性





大数据的 4V：大（Volume）

这个现象在上一章已大略解释过，人类产生的数据总量为什么呈爆炸式增长，在这里，我们以数据储存单位做更进一步的说明。根据科技研究公司 IDC 估算，仅在 2011 年一年内，就有高达 1.8ZB 的数据产生，这相当于每一个美国人每分钟发 3 条 tweet，还不停地写上 2.6976 万年的数据量；或者一个人 24 小时全天候无间断地把 2 000 亿部高清电影（每部长度为两小时）花上 4 700 万年看完的数据量。

Twitter 每天就产生 7TB 的数据、Facebook 更高达 100TB，有些公司更以每小时好几 TB 的速度迅速累积数据量，而企业内部储存系



统保存数 PB 数据的例子也比比皆是。根据麦肯锡调查，美国 17 个产业中就有 15 个产业，单个公司储放的数据量比美国国会图书馆更多。想象一下，当您在读这本书时，实际的数据量早已超过此时的预估了。

仔细想想，我们会迷失在数据汪洋中绝非偶然。当你把智能手机从皮套里拿出来看时间，这是一个“事件”（event）；当你赶着上班，捷运的门打开时，也是一个事件；进入公司前打卡、出差时在机场登机、度假出游时开车经过电子收费站、下班后上网在 iTunes 下载新歌或坐在沙发上转换电视频道，每一个动作都是一个事件，也全都会产生无数的数据资料。

图 2-3 数据存储单位

存储单位（英）	存储单位（中）	说明
字节	位元组	档案存储容量的最小单位
Kilobyte（KB）	千字节	1024 B
Megabyte（MB）	兆字节	1024 KB
Gigabyte（GB）	千兆（吉）字节	1024 MB
Terabyte（TB）	万亿（太）字节	1024 GB
Perabyte（PB）	千万亿（拍）字节	1024 TB
Exabyte（EB）	百亿亿（艾）字节	1024 PB
Zettebyte（ZB）	十万亿亿（泽）字节	1024 EB
Yottabyte	一亿亿亿（尧）字节	1024 ZB

另一方面，数据对于现代人，就如同水之于沙漠中的旅行者一样，



人们极尽一切可能地追踪和记录数据（这还不包含分析已经储存起来的数据），饥渴地囤积数据，为的就是应对多变的环境。

例如，一家远在欧洲的汽车挡风玻璃制造商，期望能预测中国汽车的增长率，以便在这个全球企业觊觎的大市场里，制订更为精确的销售和运营计划。所以他们需要收集有关中国汽车产业的数据、全世界其他的汽车产品在中国的发展、甚至是中国政府颁布的交通政策、中国13亿人口对于商品的消费反应。这些分析预测结果都得倚靠大量的、非传统的数据源才能分析。

另一个数据量大增的推动力，来自无所不在、大量增加的传感器，而且这些传感器的装设很可能是在你意想不到的地方。以美国明尼苏达州的密西西比河大桥（the I-35W Mississippi River Bridge）为例，2007年8月，这座有40年历史的4车道大桥，在交通尖峰时段突然倒塌，造成13人死亡、近百人受伤，成为美国自1983年以来非天灾或外力因素，所造成伤亡最惨重的桥梁崩塌事故，事后分析原因，可能是因为当地持续暴增的车流量，造成连接跨距和承重柱的钢梁上16个加固板受损，桥体无法承重而崩塌。

新建的圣安东尼瀑布大桥（St. Anthony Falls Bridge）采用高性能混凝土制成，提供桥体强劲的支撑力以及高度抗腐化功能，预计使用年限高达百年。此外，为了避免憾事再度发生，桥体的基底部份安装了323个传感器，用以对桥体结构进行永久性的检测，只要温度或湿度稍有变化，或是发生承载量骤然增减的异常压力点，都会立即产生数据资料，并被拿来存放、分析和判读。



现在你应该了解到，在这个处处布满传感器，每个人随时可以上网的情况下，数据量不仅大，而且是非常大！仅仅10年前，我们还得特别列出全世界有哪些企业数据仓库中心的数据储存总量超过1 TB；现在，光是个人家用的计算机硬盘容量便动辄超过1 TB，而且TB已经不够用了，现在大家谈到数据库，谈的是PB级的存放空间和计算能力，而且不可避免的是，未来还会朝向ZB级迈进。

大数据的4V：杂（Variety）

除了庞大的数据量之外，数据处理中心面临的另一个新难题是：数据种类繁杂。

传感器、智能型设备和社交科技全面渗透到日常生活和工作环境中，企业要处理的数据也越来越复杂，不仅有传统的关连式数据，还有来自网页内容、网页日志文件（包括单击流）、搜索索引、社交媒体讨论室、电子邮件、文件、主被动系统感应数据等各种各样原始、结构或半结构的数据。

而且，大多数数据无法以传统的数据库技术管理。既有的系统根本难以储存和分析这些数据，更遑论从中解读出什么意义了。尽管有些企业已经很积极地想要驾驭大数据分析，但是绝大多数的企业，现在才正开始了解大数据分析所蕴含的商机，或体会到不懂大数据分析将付出多高的代价。

今日，组织的成功却又取决于，能否从各种各样数据中过滤出有



价值的策略来辅助业务，而这些数据绝不是只局限于结构化数据。举例来说，收集 tweet 或微博消息时，你看到的是采用 JSON（Javascript Object Notation）格式的结构程序，但消息本身的文字却是非结构化的数据，而这些非结构化数据又包含了各种形式。以 Facebook 来说，用户平均每个月发表高达 300 亿则的内容，其中包含了文字、图片、影音各种不同的数据，而光是发表的相片一天就超过 30 亿张，大家最常点的“赞”则有 27 亿个，更别提难以计数的小通知、生日祝贺、邀请等。

这些影音、文字和图像，以程序编码来说，很难被传统的关系数据库所储存，或者是雨量、温度、车流量等这种随时动态变化的数据，也难以套入数据库既定的类目中保存和管理。

以载货火车这种看来再寻常不过的东西为例，为了应对安全问题，每一辆列车或每一条铁轨上，其实也都装上了数百个传感器。列车上的传感器会记录每节车箱和车内每个零件的状态，工程师可以利用车体传感器收集较易耗损零件（如轴承）的数据，在零件故障之前更换或检修；调度人员则利用每隔几米就要设置的铁轨传感器，追踪货运和物流路线的卫星定位数据，再加上平交道传感器、可能造成铁轨移动的气候等各种来源的数据，控制每班火车所承载的货运量、发车和到站时间等，这些都还没有纳入车站本身对于货车和客运车的调配数据。

由于过去曾发生过火车出轨而造成重大生命财产损失的事件，有些国家或地区政府明定此类数据必须予以保存，以预防类似的灾难发生。



很明显的，光是载货火车这一项交通工具的数据量已是非同小可，而且这些不断产生的大量原始数据，不仅存放空间有问题，而且数据格式纷杂，根本无法、也不适合以传统的关系数据库（Relational Database）处理，所以我们才需要新的分析方式。

2012 年 6 月，IBM 就利用大数据分析系统，协助医治波士顿严重的塞车问题。分析师以现有的交通数据、手机上的 GPS 定位系统和来自 Twitter 的数据为主要来源，重新拟定交通管理政策。例如安装在 iPhone 上的移动应用分析软件，就像是移动的智能仪表板，加上 GPS 系统可以实时采集的交通号志、二氧化碳传感器甚至车速的数据，可以帮助开车民众重新调整路线。

除了来自 GPS 和手机传来的数据之外，再加上 Twitter 帖子了解民众的意见和需求，这些不同来源、形式各异，每秒钟数以百万计的大量数据，经过分析处理后，波士顿政府已着手制订更好的自行车、泊车和交通管理政策，希望大幅降低城市的碳排放量和塞车情况。

以往，数据库管理人员常花上大笔时间处理世界仅 15%的数据，也就是那些格式整齐、符合既有格式的结构化数据。但事实上，全世界有 85%以上的数据都是存在于社交网络媒体、电子商务等之中的非结构化数据，或顶多是半结构化的数据。面对包罗万象的大数据，旧有的数据库已不足以应对如此庞杂的数据格式，极需找寻新的解决方案，因为这类数据虽是不断拉高数据量和数据变化速度的“元凶”，却也可能是从中分析出高价值信息的“宝藏”。



大数据的4V：快（Velocity）

数据量和种类改变了，制造数据的速度也和以往大不相同，更具体一点来说，实时变动的流动性数据（Data In-Motion）已成为大数据分析时代的另一个挑战。以爱尔兰的戈尔韦湾（Galway Bay）为例，这里和世界上许多水域都一样，正面临着水质污染、鱼群数量减少和气候变化的威胁。因为地理环境的缘故，每当漏油或其他污染事件发生，戈尔韦湾中污染扩散的速度远比公海快，因此爱尔兰海洋学会（Marine Institute Ireland）与IBM合作，在海湾中装设数百个浮标，浮标上带有传感器，通过无线电与网络链接，实时测量海洋与气候环境的变化。

通过频繁的采样和追踪，任何细微的水温、浪高、洋流状态、盐度和含氧量变动，都会被实时记录下来，科学家从不断更新的流动性数据中掌握海洋生态的变化，除了以此及早采取应对措施（例如因污染值加剧而关闭海滩）之外，也希望利用传感器测量浪高，找到不同时段里波浪发电的最佳地点。

另一个不断产生流动性数据的应用实例是一秒也无法耽搁的医疗业。以早产儿护理为例，由于早产儿出生时免疫系统尚未发育完全，加上经常需要插管、注射或做各种检查，万一生病或感染，病情变化可能会来得又快又急，特别危险。所以，加拿大安大略理工大学建立了早产儿健康监护系统，在早产儿的身上和周围装设传感器，收集传感器和其



他监测设备生成的心跳、呼吸等数据资料，每秒最多可产生高达 512 个的监测值，通过不断更新的流动性数据，协助医护人员提前 24 小时预防早产儿因败血症引发的感染。

但是，对于旧有的 IT 系统来说，这些不断更新的流动性数据，就好像是下了一场永不停歇的倾盆大雨一样。在正常气候型态下，每年降雨量多半落在一定的范围内，就算偶有强台风挟带暴雨，雨量也不易超出排水系统的设计标准，这就像是以往传统企业 IT 系统设定的吞吐量和数据库容量一样，只要流进来的数据量（降雨）不超过数据传输的最高限制，就可以被 IT 系统（排水沟）收集起来，再输送到负责存放的数据库（集水池）内。

而现在的状况就好比极端气候，在各种流信息的交融积累下，旧有的 IT 系统就像是正在经历有史以来最大规模的暴雨，这场雨下得又猛又急，让排水系统完全招架不住。然而，在分秒必争的商业环境里，对企业来说，已经没有时间容许 IT 系统好整以暇地等候数据搜集完成，再细火慢炖进行分析，尤其是在涉及医疗护理、电子制造业制程改良的数据分析，甚至得在微秒间获得结果，并在事件发生的当下立即做出决断，才能产生价值。

过去习惯处理静态数据的数据库管理模式，已显捉襟见肘，IBM 所提出的“流计算”（Streams Computing）成为解决流动性数据问题的新出路。以往，企业数据库都是以每周或每月为周期，进行批次性的数据统整，再运用这些已经整理好的结构化静态数据进行分析；数据库完全无法处理异动频繁、流入量庞大、用户需要实时响应的动态流数据



（Streaming Data）。

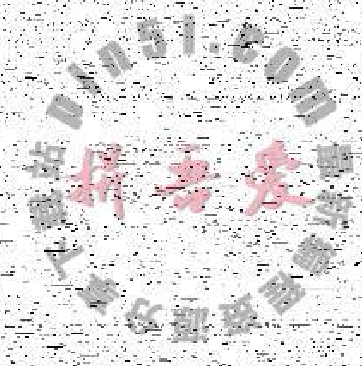
流计算系统则是利用多节点 PC 服务器的内存大批处理，无须等待数据储存，不用从企业数据库提取数据，就可自动汇集动态数据，直接处理流动的多元结构数据，其分析反应速度更可控制在微秒（micro-second）之间。也就是说，在流动性数据被储存之前，流计算系统先就其进行分析，比如银行想要导入防欺诈的风险控制机制时，只要设计好一定的逻辑后，系统会根据此逻辑先判定是否为欺诈行为，经比对后，确定交易行为，此笔事务数据才会写入数据库。

这样一来，像是智慧电表等传感器（sensor）产生的数据、网络/系统/Web 服务器/应用服务器的数据日志（Data Log）、需要高速交易的金融数据等，全部可以先分析比对，再存放到数据库中，对企业来说，可以早 1 微秒比对手发现新趋势、新问题或新机会，在混沌多变的商业环境里抢下先机。

大数据的 4V：疑（Veracity）

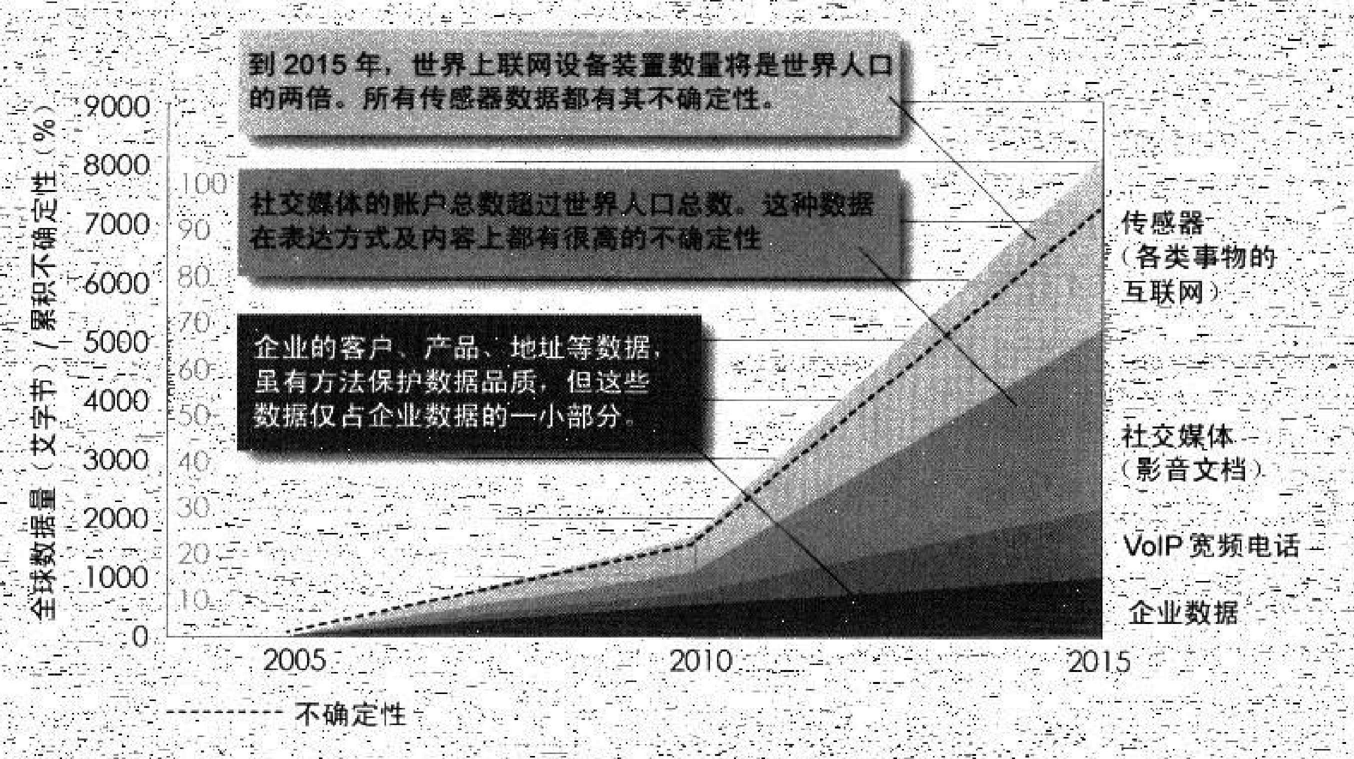
不过，在数据被处理、分析和应用前，我们应该先问一个非常关键的问题：你得到的数据可靠吗？

过去，很多数据取得后，包括企业在内的各种组织，通常会仔细查核内部数据的来源，所以数据的可靠度较高，但在网络语音通信（VoIP）、社交网站和传感器技术的蓬勃发展下，破碎的、不完整的、不可靠和真伪难辨的数据越来越多，甚至有研究单位预估，到了 2015 年



时，在全球采集的所有信息中，将有超过 8 成属于不确定可靠与否的信息（见图 2-4）。

图 2-4　2015 年，全球将有 8 成数据的可靠性不明



数据来源：IBM 汇总的 IDC 及 CISCO 数据

数据可靠性不够高，势必会影响信息分析的价值。譬如，实时整合和分析好几百万个患者的病历，可以帮助研究人员在传染病爆发大流行前及早采取应对措施，但万一同一种疾病或治疗方法在病历中有好几种不同的名称，就可能造成分析软件无法判读，导致评估和预测结果不够精准。

随着数据真实性与可靠性的问题越来越大，除了大数据的巨量性（Volume）、多样性（Variety）和实时性（Velocity）之外，第 4 个 V，也



就是数据的不确定性（Veracity）也开始受到重视。

很多因素会造成大数据的不确定性，包括“制造过程的不可靠”；很多事物的处理过程就算设计得再精密，产生的结果也总会有些难以预测或掌握之处。譬如，先进的半导体晶圆制程也无法确保产品 100%的合格率；送货路线规划得再好，也无法精准预估尖峰时间从甲地送货到乙地的车程；传感器所感测的数据也可能因为环境变化或使用年限而发生错误、甚至数据传输的过程也可能被恶意破坏，这些都可以让数据制造的过程中产生不准确的数值。

例如，曾有医疗研究单位为老年人的家中装了无数的感应监控装置，有的甚至装在地板下面，就此测知老年人的生理状况和疾病发生之间的关系。某日研究人员非常震惊地发现，一位老年妇女在就寝和早餐之间的这段时间里，体重居然增加了八磅，这是否代表她水肿的程度已达到危险等级？结果研究人员实际到老年妇女的家走一趟，才搞清楚，她在固定的时段体重增加，只不过是因为小狗习惯跳上床和她一起睡罢了。

再者是“数据内容的不可靠”，尤其是由“人”所产生的数据，不可靠的程度特别高，这主要是因为：

1. **蓄意欺骗**：网络上的信息有多少是真的？有些人在 Facebook 上发布假消息，有些人在微博上有好几个假 ID，有些博客拿钱帮美容企业写护肤产品的推荐文，有些厂商在网络论坛上发动“刷屏”以压过消费者对产品的差评……这类的例子屡见不鲜，只要荷包



够深、时间够多，任何一个人或组织都可能在网络上制造出假民意和假趋势。

2. 无心欺瞒：无心之过也可能在产生数据的源头造成不可靠的情况，譬如，现场证人看错了，而无法在警局指认出真正的杀人凶手；有时候，无心之过则发生在数据产生之后的传送阶段，例如谣言的散布。网络就曾流传着一份含有硅灵的洗发精品牌名单，这是某家厂商蓄意攻击对手而假造出来的，但当消费者接收了错误的数据后，由 A 好心传给 B 参考，B 又传给 C，C 再传给 D……就像曾参杀人的故事一样，到最后已经真假难分，连曾参的母亲都不再相信儿子的清白。

3. 时序错乱：相信很多人都收过类似的转寄邮件，一个父亲焦急地请大家发挥爱心，协助寻找走失的 8 岁女儿，最后还留下联系电话，网友收到信件好心转发，但其实小妹妹早在走失两个小时内就自己回家，但这封在 2005 年发出去的寻人信件，却在网络流传长达 6 年，已经 14 岁的女孩通过电视新闻的镜头感谢网友，也请大家不要再找她了。当初发信的女孩父亲说，6 年来，家人每天平均接到 20 多个电话，最远还曾接到从中国香港地区、越南和乌克兰打来的，虽然网络协寻力量大，但也希望网友们不要再转寄了。这就是因为时序错乱而导致不准确的数据。

还有“分析结果的不可靠”，即使是由全世界最聪明的数学家设计出



最顶尖的演算系统、尽可能收集到最完整的数据，想要在大、杂、快的大数据中，整理出一些可理解的头绪，都必须归结出众多现象的“最大公约数”。换言之，所有的演算模型都只能估算出“近似值”，而不是绝对的真实，这也是为什么气象预报系统虽然越来越精密，预报结果仍难免不准。

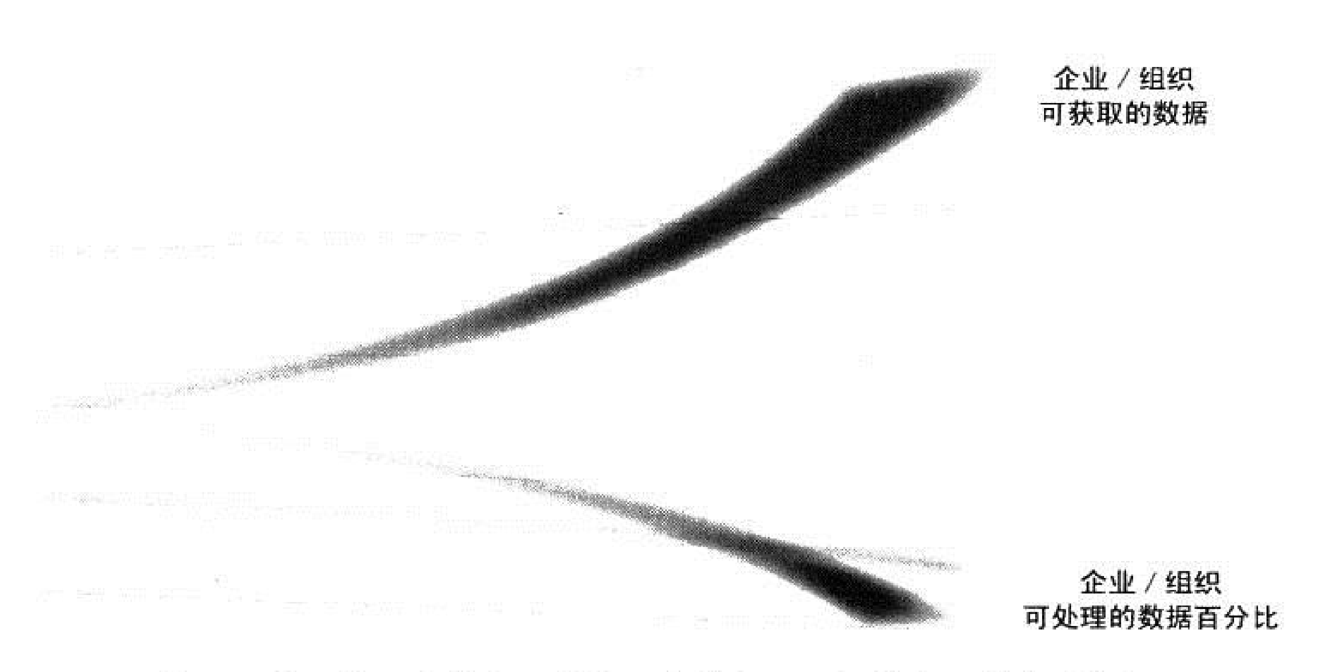
因此，当我们试图收集大量的数据，希望从中找出某种规律和趋势时，也必须思考这些庞杂的数据量中有哪些不确定因素存在，否则不可靠的数据会形成不确定的分析结果，进而影响后续决策的价值。

数据里的含金量

综上所述，我们可以知道在现在这个信息泛滥的世界中，排山倒海而来的数据，大多是以最原始的、非结构化或半结构化的形式存放，导致很多组织根本不知该如何从中提取有价值的信息，甚至也无从判断有哪些信息值得保存。

如“大数据分析”一词所隐含的意思，当今组织面对的是极大量、极庞杂，且不断变动又难辨真伪的数据。不懂得管理这些数据的组织，一不小心就会被数据汪洋所淹没。然而，与此同时，各产业的企业也陷入两难。随着数据量持续增加，企业所能处理、理解和分析的比例却迅速下滑，因而形成了莫名的“浑沌区”（见图 2-5）。浑沌区里藏有什么？无从知晓。或许是货真价实的稀世珍宝、也可能是毫无价值的弃土废矿，但是，最痛苦的就是“不知道”。

图 2-5　数据量持续攀升，但企业/组织可分析的比例却迅速下降



最近 IBM 有一份研究指出，虽然现在的企业已经有办法储存所有想要保存的数据，但因为产生数据的方式与速度前所未见，有半数以上的企业领袖发现自己无法借此取得经营事业所需的商业洞见。在这样的环境下，一场庞杂的信息混战正让企业陷入进退维谷的困境：取得业务洞察力的渠道与平台变多了，但是，随着数据的矿藏量越堆越高，企业却有如在滚滚河流中筛取砂金，要在最短时间内分辨筛网中的是泥沙还是金沙，其实非常困难，因此从信息洪流中所能炼出的真金比例也越来越低，而且下降的速度越来越快。

但是，只要有正确的科技平台、使用正确的方法来分析各项数据（至少是有用的数据），浩瀚的数据原矿也可能化成开启商机、了解客户和纵横市场的密钥。



我们借用采矿的比喻为本章作总结。很久很久以前，矿工是先“看到”埋在土石中闪着金澄色泽的矿块后才开始挖掘，然后再设法在发现第一块金的地点附近找寻更多可开采的金矿。随着掘金的矿工越来越多，这个区域的矿藏越来越少，虽然外围还有更大的矿脉，或许就在旁边的山上、又或许在好几公里以外，但是，人们光用肉眼是无法找到新矿藏的，自此，采金就变成了一场赌博，梦想着一夕致富的工人，只能在上次采出金矿的地方执着地挖掘，却没人知道找不找得到金子。

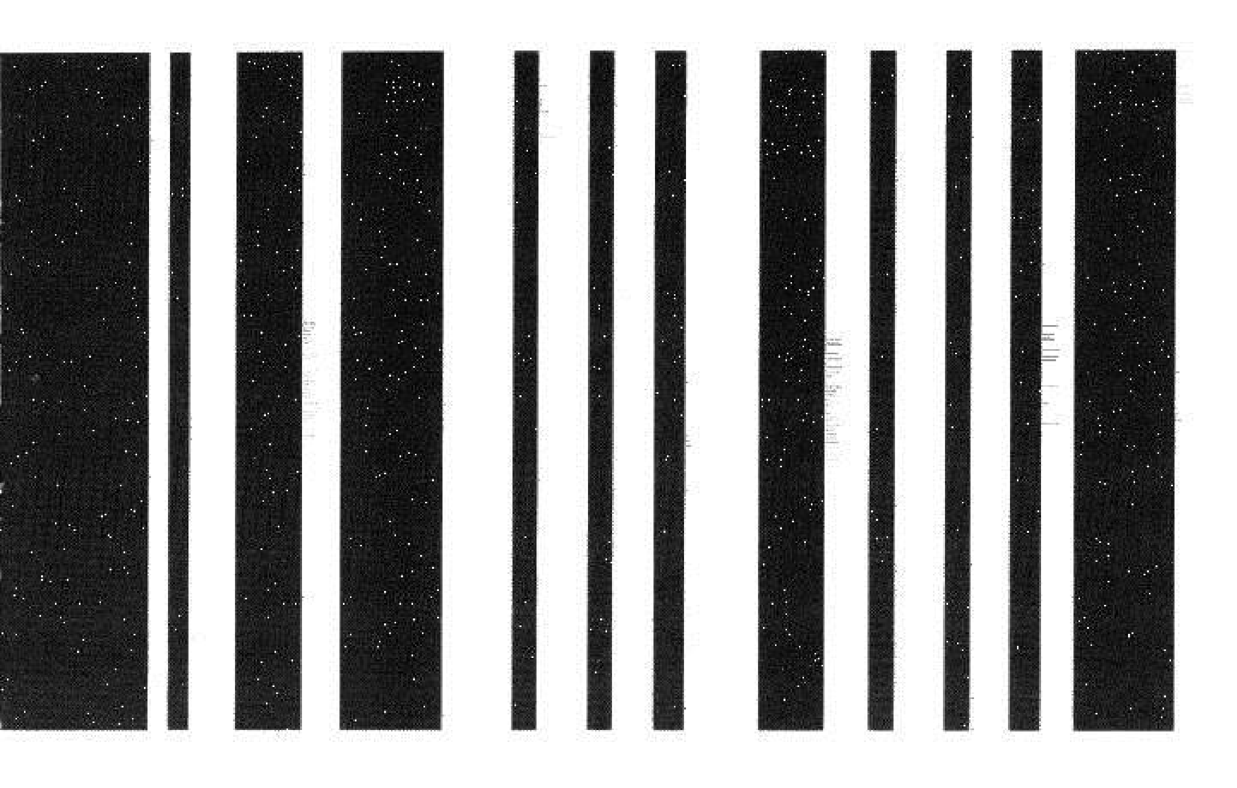
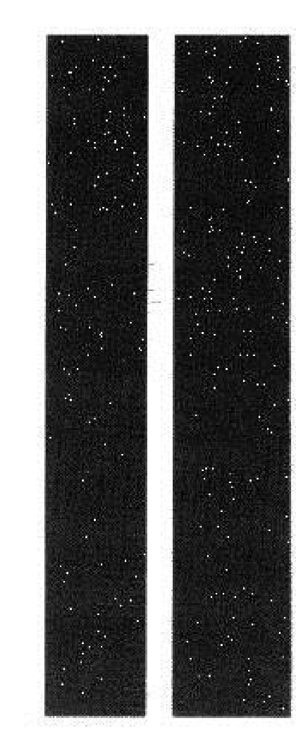
一般而言，矿体的金矿品位（ore grades）每千克至少要有30毫克（mg）的黄金含量，肉眼才能看得到，但目前全世界绝大多数的金矿都已是肉眼看不到的，导致大多数的真金（高价值数据）还深埋在废土砂（低价值数据）里。但是，今天的采金术不一样了，矿业公司豪掷资本、买下昂贵的探勘器具，这些机械一次可探勘几百万吨的土石。

这说明了只要用对工具，就能用符合经济效益的方式筛掉砂石、找出闪亮的金屑。把这些金屑收集起来，镕铸之后便能变成价格高昂的金条。挖掘大数据也是一样。过去，数据量太大、毫无价值的数据太多、“赌博”的代价太高，没有一家企业有本钱再用过去既有的流程——过滤所有的数据，只有利用新的方法和工具，才能以最经济的方式储存和处理数据，并从中挖出值得锻铸与收藏的珍宝。



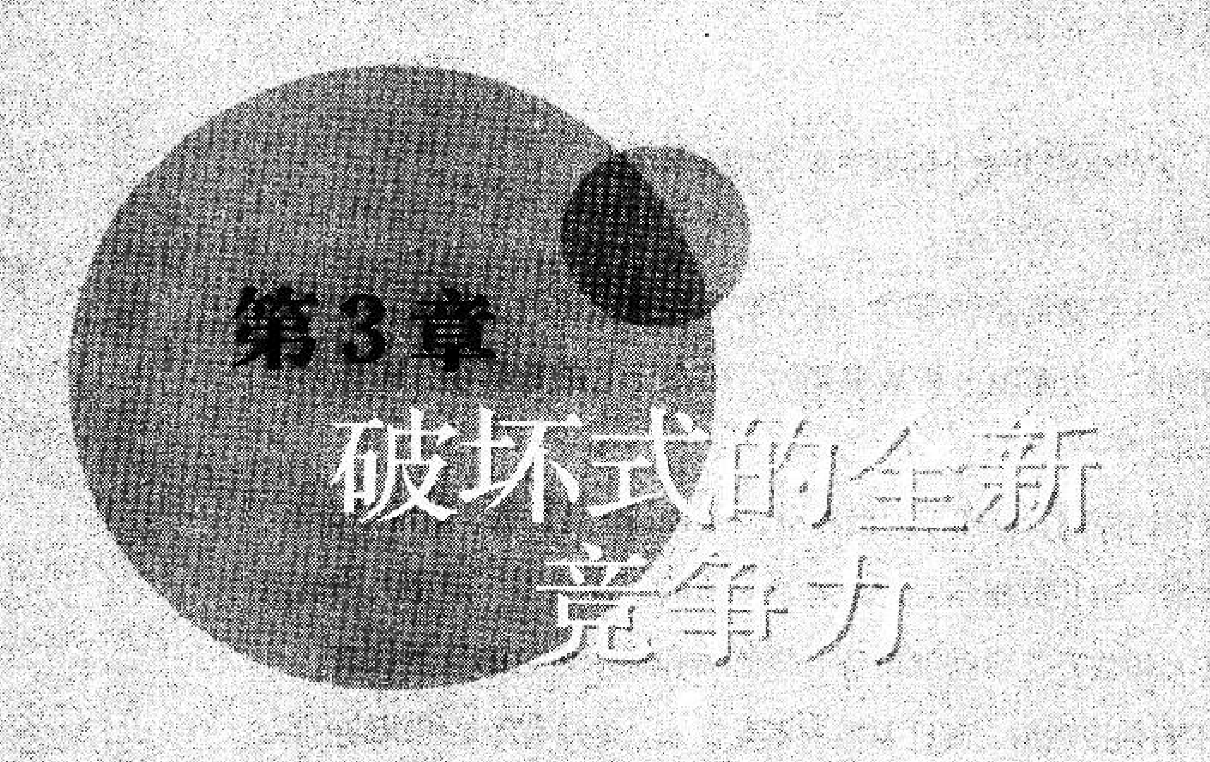
Part 2

大数据大商机







第3章 破坏式的全新竞争力

象你正在参加一个益智节目，电视墙上浮现的题目是：在大数据分析中下列哪一部分是最重要的？

1. “大”的部分（the“big”part）
2. “数据”的部分（the“data”part）
3. 两者都是（both）
4. 两者皆非（neither）

请你花30秒来思考这个问题，然后锁定答案。

正确答案揭晓！答案是4。

为什么？这是脑筋急转弯吗？或许有人会这么抗议。其实，这个小玩笑要强调的是，大数据分析最重要的，不在于它的数据量，而是你可以用它来做什么！

前两章我们提到，在现今的世界中，Facebook是数据、微博是数据、对话也是数据；这些数据不一定是有意创造的，甚至不一定是人为创造的：几十亿人和数万亿台智慧设备、传感器和形形色色的仪器仪表，一天24小时不断创造数据，而且80%的新数据完全是没有经过组织的内容。

当我们感知化的程度越高，使用的传感器也越多，这些从不同渠道所获得的大量数据，如果真要进行实时处理和分析，估计必须以每秒6万多次的操作速度采取行动，而这个速度比蜂鸟的翅膀扇动速度还快上300倍。

对于企业或组织来说，大数据分析的意义并不是强调我们有多容易



在庞杂的数据堆栈中迷失，或是要实时分析这些数据有多困难，而是要让企业或组织知道，当大数据大潮来袭，如果只是消极地扩充储存空间，却无法分析大数据的使用意义，这些数据就相当于堆放在仓库里的大型垃圾。

反之，如果这些在传统的商业智能（BI，Business Intelligence）解决方案中可能尚未被开发，也可能因为太过杂乱而被弃置的大量数据，经过交织、整合在一起后，它的内在价值可以让公司增加 50%的新客户，让政府减少 30%的成本，它就不只是一份大而无用的数据，而是一项宝贵的竞争优势。

把从不同渠道获得的大量数据，转变为经过组织的信息、甚至知识，这就像是一个人同时拥有听觉、视觉、味觉和触觉一样，将这些“感觉”综合起来，才能感知到我们身处的每一个瞬间，然后应对各种状况，做出适当的反应。而且，通常你所能接收到感觉越多，对于事物的感知能力也越强，所以可以连接的不同种类数据越多，你获悉的洞察力（Uncovering Insights）也就越细致、准确，进而更容易做出实时的决策。

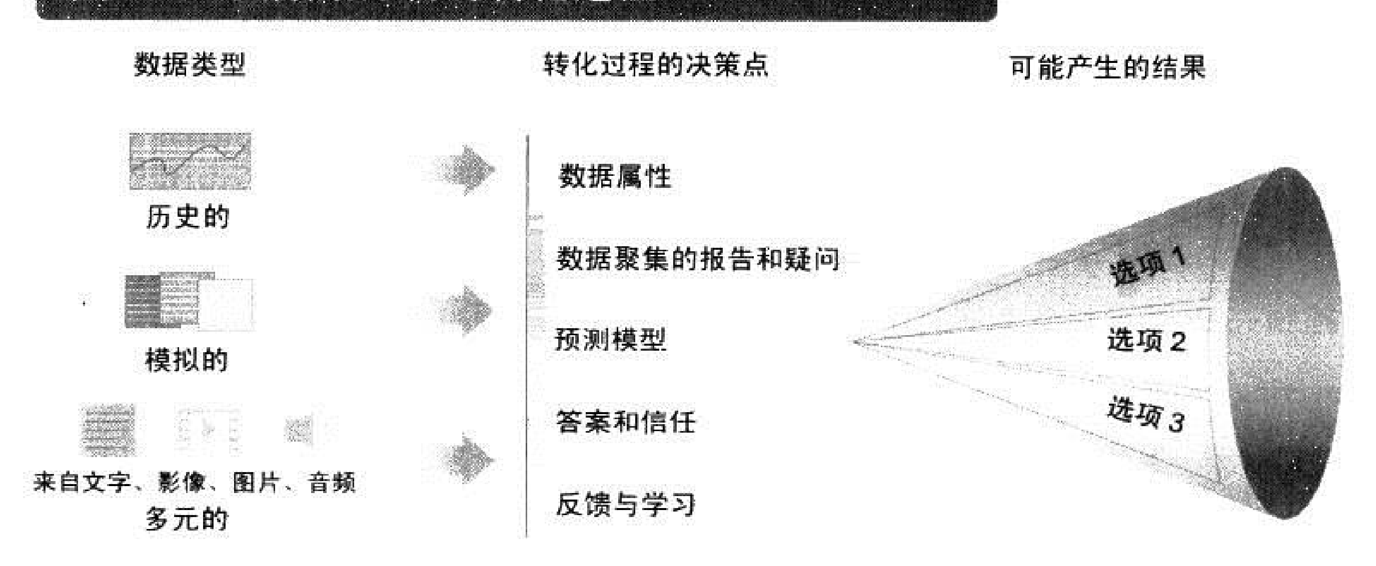
例如欧洲一家连锁烘焙坊与 IBM 合作，就从人潮流量、客户信息、气象信息中洞悉一个现象——每当下雨天，女性消费者喜欢吃蛋糕；晴天则是潜艇堡、三明治销量较佳。以往企业只能凭经验发掘这种现象，但现在有了数据资料左证，便可以利用整合每日气象、原物料、生产、销售等大量数据后进行预测，促使中央厨房调整出最佳产品组合，这家连锁烘培坊因此提升了 20%的获利。

因此，我们在讨论大数据分析时，它真正的价值是在于此趋势背后

所包含的“分析学”（Analytics），也就是以系统化的方式，把观察和经验转化成知识，让企业和组织在应对各领域数据的快速增长时，除了储存之外，能够更进一步地分析数据、提取信息、萃取知识，并且应用在决策辅助上。

从先前的内容中，我们可以了解到“大数据＝大量+复杂+快速变动的数据”，而大数据分析（Big Analytics）正是将这些来自历史的、模拟的、多元的、正在产生的庞杂数据，转化成有价值的洞见，进而成为企业或组织决策辅助的选项，如图3-1所示。

图3-1　大数据分析的转化过程



简而言之，大数据分析就是化繁为简、化数为宝，让企业和组织可以结合分析结果和采取行动，做出更好的业务决策。以零售业为例，企业可以借由收集消费者在店内的走动实况，以及与商品的互动影像，配合大量的事务历史记录数据，重组商品销售的类型、位置和调整售价的时间，目前已有卖场因此减少了17%的存货，并增加了高利润率的自有



品牌商品销售比例。

我们也可以利用大数据分析打造实时、个性化的服务，尤其是可以将消费者细分成更微小群体的网络购物。现在已有网购企业通过网络点击流，追踪个体消费者的行为、更新其偏好、并实时模拟后续的购买倾向。这种实时性的精准营销，不仅可预测客户再次光顾的时间，同时也可以针对个人需求，促使最有价值的客户购买高利润率的商品。

除了零售业之外，这种量身订做的数据应用服务，其他行业也能从中受益。例如，一家寿险公司依照客户风险、财富变化、家庭资产价值和其他输入数据，对每一名客户提供量身订制的保单。

至于制造商，他们已经利用大数据分析系统，以来自生产线的传感器数据，建构自动调节的流程以减少制造过程中的损失，同时避免成本高昂（有时十分具有危险性）的人工干预以增加产出。在目前最先进的数字化油田中，仪表不时读取有关井口状况、信道和机械系统等各类流动数据，并自动将结果输入实时分析中心，以调整油量生产速度和停机时间，一家大型石油公司因此减少了10%～25%的运营成本和员工成本，产量提高了5%。

汽车制造业则是试图打破藩篱，寻求上游供货商和下游销售点的外部信息，以大数据分析整合来自不同系统的数据，以便在新车款的设计时间三方合作共同开发，此举不仅控制了决定最终制造成本的关键因素，同时也更符合市场需要。

对于内部管理来说，大数据分析可以利用对照实验，测试各种假设和分析结果，从而做为投资决策的指标，以及改善财务表现和产品性能，



就像是许多金融机构用来做投资仿真的数学模型。而现在，包括像是麦当劳等大型零售业，也开始在每一个地区的部分门市，安装了搜集营运数据的传感器或监控装置，用来追踪店员与客户互动的流程、客流量的模式，以进行菜单变化、餐厅设计以及服务流程的改善。

科技和商业的新变革

大数据分析可以协助企业和组织有效提升竞争力，但是对很多人来说，目前这个概念还处在初期的萌芽阶段，因此业界也产生了大数据分析到底是变革（revolution），还是进化（evolution）的争论。

有人认为数据分析技术早已存在多年，最古老的名称叫做“统计学”（statistics），新潮一点的说法可能叫做“数据挖掘”（data mining）、“数据仓库”（data warehouse），不管名词如何转变，基本上都是收集、整理、分析与诠释数据。那么，大数据分析不过就是一种数据可处理量更大、速度更快的高级技术而已，它和以往的数据分析又有什么不同呢？

在回答这个问题之前，让我们先回到以往的数据分析方式上。一直以来，IT 人员在处理数据分析时，都是在寻求所谓的优化（optimizing）解决方案，也就是在一个有许多限制和条件相互冲突的环境下，找寻一个最适合方案的过程。最合适的答案代表最好的妥协，也就是为了让计算机计算量可以负担得起，必须牺牲掉很多他们认为不需要或没有意义的数据。

换言之，优化其实就是简化（simplify）！但是，这种传统的数据分析



方式无形中也把优化的目标降低了，就像是我们只看我们想看的，只知道我们可以知道的，而这种以管窥天的做法，不仅难以知道事情的全貌，同时也限制了发展的境界。大数据分析的概念则颠覆了这种想法，IT 人员第一次考虑的是如何收集更多的数据来分析，而不是怎么减少数据量，研究人员不是要舍弃数据、简化模型让计算机可以计算，而是要想办法收集以前因为计算机无法计算而舍弃的数据，让计算分析出来的结果价值更高。

在系统设计上的思维也不一样。在网络、智能手机、传感器兴起的这 10 年内，因为 IT 设备成本一直下降，面对大数据，大家谈的都是如何用更便宜、空间更大的方式储存数据，但这种想法还是只能先把数据储存起来。尤其是在传统的数据库中，数据的关联性结构必须像是图书馆的索书号，依照固定的格式编码，如果数据要更改，或者要将 A 渠道传送而来的数据和 B 渠道及 C 数据库的数据结合，IT 人员就得大幅更改数据库的整体结构，因此所有存放的数据还是只能用来做事后的分析。

但是现在，大数据分析的架构不同了！以前进行数据分析时，IT 人员是以程序为核心，把散落在外面的数据，放到储存系统的结构里，然后拿进内存，交由计算机程序计算，所以数据得经过提取、处理、分析，最后等上数天、数周的时间才能得到结果。大数据分析的概念则是以数据当作核心，外面放了很多的程序和硬件，依据不同的需要对应不同的处理程序，产出个性化、以毫秒为单位，并可随着时间推移的实时分析结果。

毫无疑问，这些想法和概念和以往 IT 人员所学的数据库概念完全不



同，也将会在未来十年内，促使科技界的软硬件研发产生很大的变革，而不只是进化。

除了科技界之外，大数据分析也正在形塑一个崭新的商业运作规则。30 年前我们要完成一个人的传记，作者必须像个侦探一样，爬进你家的阁楼或仓库里，翻阅成堆的信件、照片、电话账单，同时采访你的朋友、邻居与同事，才能拼凑出一个人大致的轮廓。

但是现在，当人们生活的历史记录已从原本大多储存在纸本与模糊的记忆中，改变成以数字档案传送，不论是文字、音乐和影像，甚至人们的行为都逐渐转化为可记录的 0 与 1 之后，现在从你的电子邮件、数码相片，以及和朋友的 MSN 对话、Facebook 留言，就可以得知你的讲话方式、行事作风。

也因此，早在十多年前，企业就开始想办法收集、分析和运用这些被位化的个人数据，并期望能以此更准确地针对不同目标族群营销商品，或者是利用它来开发新的消费商机，例如大家非常熟悉的 CRM（Customer Relationship Management，客户关系管理）。

CRM 系统利用累积消费者的行为数据后加以统计分析，继而解读人们的欲望和需求，然后更精准地推销商品，就像是第 1 章提过的沃尔玛的尿布与啤酒传奇。

通常，CRM 是先将顾客的基本数据及互动历史储存在数据库中，再根据数据库的内容，针对不同区隔市场发展营销组合，提供为顾客量身订制的服务，并增加顾客满意度与忠诚度。例如，即使你不是名人，只要固定搭乘某一家航空公司的飞机，几次之后一旦上机，就会有空服员



亲切地问候你："某先生您好，有什么需要帮忙的吗？"空服员会知道常客的基本数据，就是拜 CRM 系统之赐。

虽然以前企业得面临收集数千个、数万个，甚至上千万个客户数据的挑战，但是这还可以从已知客户的消费习惯里找数据分析。时至今日，企业要面对的情况却是每一天上亿、上万亿，而且不确定从何而来的线索。以网络广告为例，每个月对台湾 Yahoo!奇摩的不重复访问的网友超过 1300 万人，占台湾地区人口的 1/2 以上，而且任何人只要造访过雅虎广告商的网站，平均会留下 2520 个线索，这也就是说，你得分析不知道来自何人的 328 亿笔细节数据，才能得知一则网络广告横幅（banner）的效果。

而且，这些数据绝大部分都是数字垃圾，如果无法下达精确的计算机指令，大量的垃圾数据很容易让公司服务器得花上 3 天来扫描，甚至是直接压垮公司的服务器的计算能力。有鉴于此，开始有人倡导数据挖掘的重要性，让企业或组织从大量资料中，发掘出潜藏的有用信息，以提供决策人员参考。

以前，从瞎子摸象到事后诸葛亮

数据挖掘是一种从大量数据中，由计算机自动选取重要的、潜在有用的数据类型或知识的过程，借由统计、机器学习（Machine Learning）、模式识别（Pattern Recognition）等技术，辅助人们进行决策判断。例如从大量的网络事务数据中，发掘顾客的消费习性和产品销售的关联性；



从以往的消费及缴费数据中，预警信用卡呆账的可能等。

但是，这一切都得基于这些数据必须有足够的“数据质量”（Data Quality），否则数据挖掘根本无法进行！

通常，数据挖掘需要4个步骤，首先是定义问题（define problem），例如某银行要推出一项新金融商品，你必须先定义需要解决的问题为“借由数据挖掘找出可能购买这项商品的潜在客户名单”。其次是数据选择（data selection），也就是从数据库中挑选数个客户基本数据的字段（假设为性别、年龄、教育程度、职业、月收入），作为数据集市（data mart），再以这个数据集市为基础进行处理。

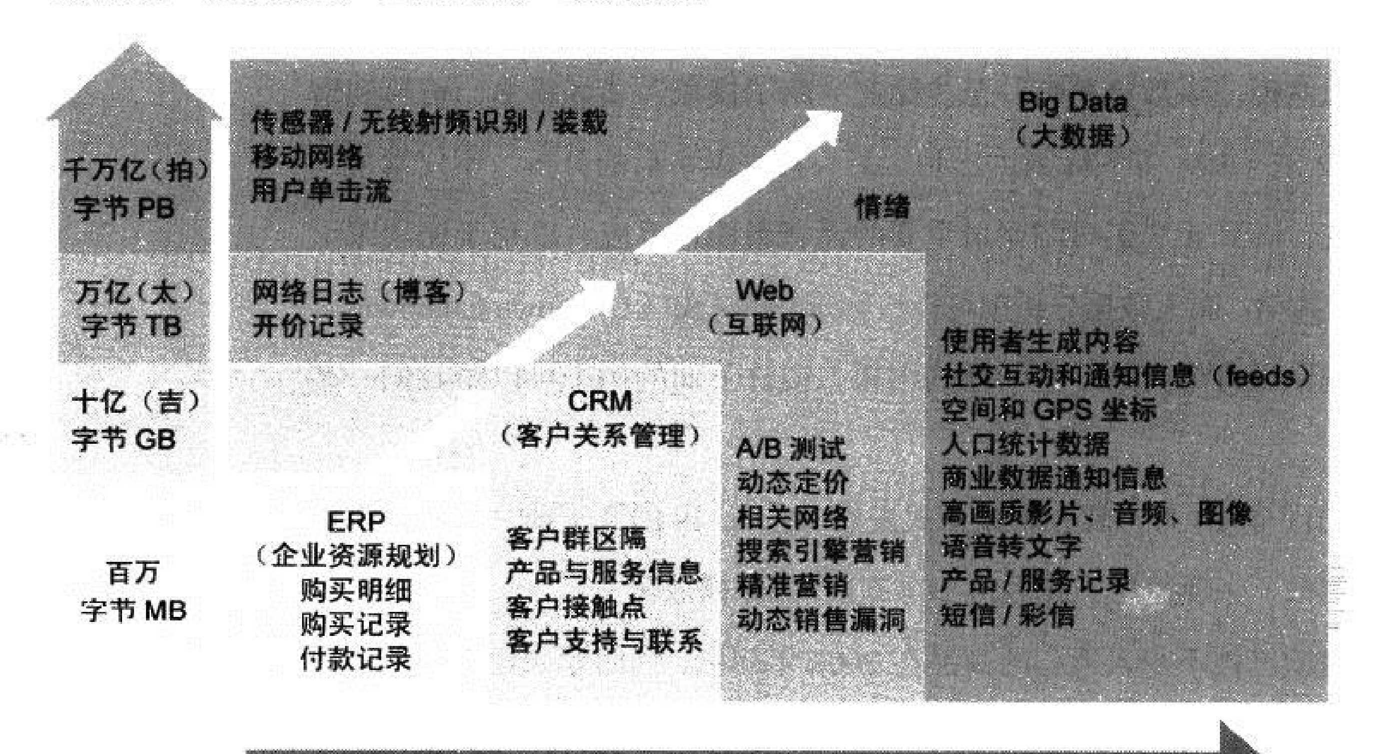
第3步是数据准备（data preparation），也就是将数据转成适当格式。转换的原因主要是你想要知道的答案可能在原始数据中并不存在，例如，一般数据中会存有客户交易的日期（Date），但你希望得知的是这3年期间客户的交易情况，此时就需要经过数据转换。另外，原始数据的缺漏、异常也都在此阶段删除或清洗，以消除不必要的偏差（bias）。

经过上述的层层关卡，你所获得的数据才算是够资格进行分析，来到最后一个步骤——知识提取（knowledge extraction），将处理过的数据经过统计、计算后，以图表、规则、数值等方式来呈现，交给该领域的专家解读。

也就是说，以往我们在谈论到商业智能（BI）时，都是通过在线分析处理（OLAP）、数据挖掘或数据仓库等技术，针对来自于ERP、CRM、SCM等应用系统的结构化数据做分析（见图3-2），即使是网络上的数据也必须先整理成同样的格式，才有办法进行分析，更遑论现在每天大量生产的半结构化或非结构化数据了。



图 3-2 大数据新时代，企业处理的数据类型



数据来源：与 Teradata.Inc. 共同整理

除了数据库系统的限制，只能分析已经在数据库中储存好、整理好的结构化数据之外，对企业和组织来说，它们希望从这些营运数据中抽丝剥茧，进而优化企业的经营体质，但是这些数据资料统统都是木已成舟的事实，即使经过整理、分析，最后还是得凭借着个人经验判断来拟订未来的发展策略。换言之，企业习惯的数据处理方式是以“事后的分析来做事前的预测”！

也就是说，以往不管是在线分析处理、数据挖掘或数据仓库等技术，



可以为像是“某某人本年度的销售业绩是多少、何时是销售高峰期”这类的问题，从结构化数据中找到很精确的答案，也就是分析“已知中的未知”（Known unknowns），在我们知道的问题中去找答案。

虽然说“事后诸葛亮”至少比起“瞎子摸象”要好得多，但等到借由 ERP 或 CRM 等实际数据出炉之后，再来做事后的分析，不仅常常为时已晚，而且无法实时洞察最重要的消费者使用反应，以至于无法跟上市场的脉动。而大数据分析则可以给你未知的未知（Unknown unknowns），也就是你没有想到的一些问题的结果，或许下面的例子可以给我们一些省思。

2011 年 10 月，就在苹果正式发布 iPhone 4S 的几天前，微软非常低调地发布了一项消息：此后将不再制造 Zune 播放器。

这款已被不少人遗忘的商品在 2006 年问世，当时苹果已在全球卖出超过 1 亿台的 iPod，而微软认为便携式多媒体播放器市场处于婴儿期，增长空间还很大，因而投入庞大的预算开发硬件，并与世界四大唱片公司达成授权协议，预备和苹果计算机在多媒体播放器和移动音乐下载的市场中一较长短。

虽然，当时 Zune 在美国的市占率仅 9%，却还是仅次于 iPod 名列市场第二，但是第 2 年 Zune 的市占率却只剩下不到 1%，连市场前五强都进不去。

身为一个失败之作，Zune 自身的功能并非不优秀，它不仅屏幕较大，有 Wi-Fi 可具备通信能力，微软甚至还为了它花大钱取得好莱坞电影公司的授权许可，可以在上面播放好莱坞的电影以及电视节目。

一开始，基于详细的市场调查和数据分析，微软对功能强大的Zune信心十足，当年比尔盖茨讽刺苹果的名言是，“Zune的用户会喜欢苹果产品，因为苹果为数字音乐开辟了一个广阔的市场，而苹果使用者会喜欢Zune，则是因为他们人人都有个iPod!”

然而，微软没想到的是，对消费者来说，购买iPod的理由大多只是因为一种前卫时尚的品味，而不是功能。例如，微软从上到下都难以理解，iPod为什么需要搞得五颜六色、像果冻一样，结论是因为iPod功能性不足，只好哗众取宠。所以，微软决定功能强大的Zune不需颜色来衬托，只需要基本的黑、白、棕3色即可，结果事实证明，在消费者的心目中，颜色就是比功能“受宠”。

Zune失败的另一个重要原因，是在智能手机兴起，手机就能充当多媒体播放器之后，消费者更无法为只有黑、白、棕3种颜色的Zune，找到一个值得购买的理由。

这个被视为微软有史以来最失败的产品，其实代表的就是企业当前所面临的挑战，消费者在意的颜色、品味、感觉，企业都无法在既有的结构化数据中找到答案。唯一的方式是从全球各个网站上的讨论区内容、Twitter或Facebook上的帖子，也就是量更大、更新速度更快的非结构化数据中，得知消费者的反应和喜好的转变。

尤其是随着移动设备、社交媒体等新兴应用的窜红，人们“黏”在这些应用上的时间，比起从前单纯利用计算机进行文字处理大幅延长。更值得注意的是，人们不再隐藏内心的独白，转而愿意对社交好友吐露真话、分享消息。这些隐藏在非结构化数据里的真实意向，可能从少数



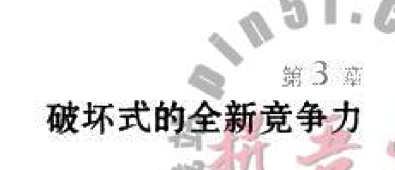
几个人的想法，迅速蔓延成为众多乡民的集体共识，进而对商品采购、品牌好恶，造成深远而巨大的影响。

现在，从测知情绪到开发新市场

在社交媒体成为人们情绪偏好的反射镜之际，企业或组织开始试图拆解这些每天大量生成的非结构化数据，期望从中挖掘更多的蛛丝马迹，更精准地测知人们的好恶。

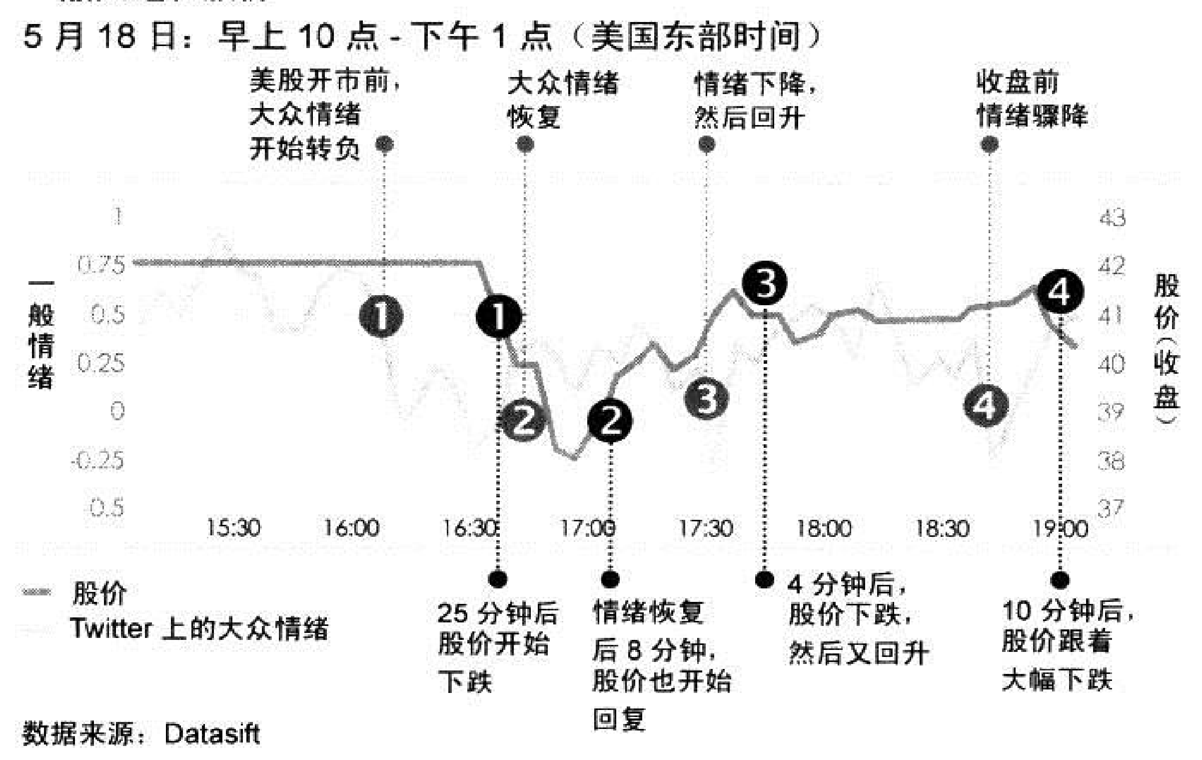
以往，不管是消费、选举、股票涨跌，这些和“民意”息息相关的社会行为，都因为没有技术或数据可以对人类情感进行量化测量，而无法得知人们如何在生活、情感、习惯的交互影响下产生行为模式的变动。企业或组织大多也只能视民意如流水，随之载浮载沉。现在，借由大数据分析已经可以对人们进行行为和情绪的细节化测量。以 2012 年股票首次公开发行（IPO）的 Facebook 股价为例，在此之前因为 Facebook 用户快速增加，导致股价预估值不断调高，没人有把握预测 Facebook 上市当天股价的走势。

Twitter 实时分析平台 DataSift 利用 58 665 位用户产生的 95 019 条 tweet，并与 Facebook IPO 当天的股价对应，发现 Facebook IPO 当天，Twitter 上发布有关 Facebook 的相关消息，其情感倾向可以做为股价走势的领先指标。以 Facebook 挂牌当天（2012 年 5 月 18 日）的股价为准，开盘后 Twitter 上的帖子情感转向负面时，25 分钟之后 Facebook 股价开始下跌；而当 Twitter 上的情感转向正面时，Facebook 股价也在 8 分钟之



后开始回弹，分析结果显示 Twitter 上每一次正负面情感的转向都可以预知 Facebook 股价的波动（见图 3-3）。

图 3-3　Twitter 上的公众情感转向与 Facebook 骨架波动成正比



发现 tweet 与股价变动关联性的还有英国伦敦基金公司 Derwent Capital Markets。该公司在 2012 年 5 月推出世界第一只基于 Twitter 上的公众情绪来进行投资的对冲基金，并承诺每年 15%～20%的高回报率。

这套社交媒体情绪追踪系统，将 1 亿条 tweet，比较正面评价和负面评价，并区分成：冷静、警惕、确信、重要、和善、快乐 6 种情绪。先前已以此预测道琼指数走势，准确率高达 87.6%，而在 2011 年 7 月



的全球股灾之中，Derwent 也维持 1.85%报酬率，领先 S&P 500 指数（下跌 2.2%）。

除了预测大盘走势之外，也有学者利用 Twitter 中包含美股代码和美元符号，针对提到标普 100 指数成分股公司的 25 万条 tweet，依照买入、持有或卖出 3 种信号进行拣选。研究结果不仅发现了 Twitter 情绪与股票价格之间的高度相关，最令人兴奋的是，依据研究者设定的 Twitter 买入信号，买入强度最大的前 3 只股票，卖空处于底部的 3 只股票，半年时间可获得高达 15%的报酬率。

不过，研究中也发现来自于 tweet 的信号过度集中也稍纵即逝，最佳的持股期限仅为 1 天，而且如果交易的股票数量多于分别处于顶端和底部的 6 只股票的话，报酬率就会降低。

虽然这些研究结果和相关投资产品目前还是受到多方面的质疑，例如有人认为 Twitter 情绪指标无法预测出给市场带来冲击的突发事件，而且许多词汇在特定行业中的意思另有所指，可能并不包含价值判断。例如 vice（恶行）这个字在金融业中通常指的是 vice president（副总裁），而 crude（残忍）则是指 crude oil（原油）。但是，社交网站上的帖子的确实时反应了人们的情绪，而人们的情绪又带动了他们的投资或购买行为，甚至形成一种意见气候，影响力扩及群体，这代表了社交媒体已成为观察大量人群实时行为的一扇窗。而借由大数据分析针对社交媒体上汇集的数万条消息，进行情感语义分析，已使得人们行为和情绪的细节化测量成为事实。

史隆管理学院教授布尔约尔松（Erik Brynjolfsson）以“现代版



的显微镜革命”来形容大数据分析的潜在影响力。显微镜是在4个世纪以前发明的，这项发明让人们看到了小到肉眼看不见的生物细胞，并可以对其进行测量，成为测量领域中的一场革命，也自此改变了世界。

大数据分析就像是现代世界的显微镜，可以测量和分析无数从传感器里流出，或是上亿条社交网站帖子的“纳米数据”（nano data），至此之后所有的决策行为都将可基于数据和分析做出，而非基于经验和直觉。

运用大数据分析社交网站的帖子结构，也发现了人际网络运作方式的另一种角度。在20世纪的60年代，哈佛大学利用包裹作为研究媒介，进行了一项与社交网络相关的著名实验，也就是先将包裹寄往美国中西部地区的志愿者，指导他们如何将包裹带给波士顿的陌生人，但却必须以邮寄的方式来交付包裹，也就是先把包裹寄给某一个你所认识的A，再寄给A所认识的B，最后送到实验指定的目标者。

结果发现，一个包裹换手的平均次数仅为6次左右，这就是所谓“小世界现象”，并以此发展了人际网络中经典的六度分隔（six degrees of separation）理论。社交媒体进一步实现了这个理论，甚至将分隔缩小为3度、4度。不过，我们进行大数据分析之后，却发现你认识但不经常联系的人，也就是在社会学中被称为“弱连结”（weak ties）的人，反而是职务空缺、小道消息的最佳来源。

原因是与关系亲密的朋友相比，这些人在和你略有不同的社交世界中生活，因此能看到你和你的好朋友们所无法看到的机会。这种以“相



近性”取代“相似性”的观察，已经成为企业开发新消费群族的重要新依据，而这类从大数据中发展出的无限商机，也成为全球大型企业探索新需求、开发新市场的关注焦点。

转变中的企业核心资产

大数据分析除了让企业更贴近消费者，并以此开发新市场、新服务之外，也正在促使传统的商业价值转变。过去，衡量一个企业最重要的资产无外乎土地、流动资金和人才等几个生产要素，如今，“数据”已成为企业另一项重要的核心资产。

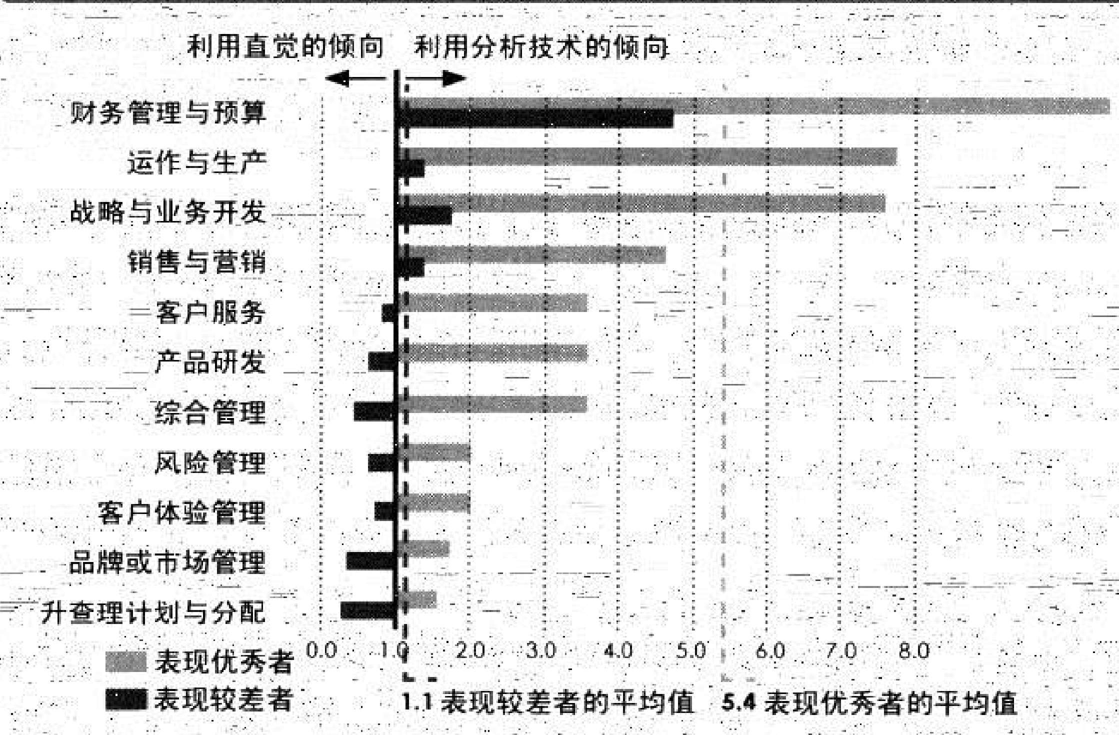
2012 年初的瑞士达沃斯论坛上，发表了一份名为《大数据，无限影响》（Big Data，Big Impact）的报告，其内容指出数据将是未来新兴的经济资产，其地位与重要性等同于黄金和货币。

除了“钱”，以往“人”也被视为企业另一项核心竞争力，因为企业运作的智慧，分布、存储在这些人才的大脑中，而在传统的商业价值中，数据的效益仅止于经过图表、规则、数值等方式呈现，交由这些管理人才或策略专家解读，再凭借着专家个人的经验判断来拟订未来的发展策略。

然而，在数据分析技术的演进下，“数据”不再是图表，而是一种能让业务流程智能运转的能力、一个让企业增长转型的关键。2010 年，麻省理工史隆管理学院与 IBM，针对 100 多个国家、30 多个行业、近 3000 位高层主管的一项调查中发现，绩效表现优秀的企业运用分析技术（5.4）比

表现较差的企业（1.1）高出 4 倍，而且表现优秀的企业全面运用了分析技术（见图 3-4）。例如，通过行为分析让客户群体持续增长；通过信息管理、业务分析、内容管理等帮助企业优化业务流程；通过财务规划分析改善企业利润结构和成本来源；通过分析洞察预测未来的法规要求及管理风险。

图 3-4　表现优秀的企业与表现较差的企业应用分析技术的倾向



注：此调查结果为由受访者指出其企业在以上活动中利用分析技术的情况。10 分表示充分利用分析技术，而 0 分表示利用直觉或非分析方法。

数据来源：MIT 史隆管理评论与 IBM 商业价值研究院分析研究合作项目（Massachusetts Institute of Technology 2010）

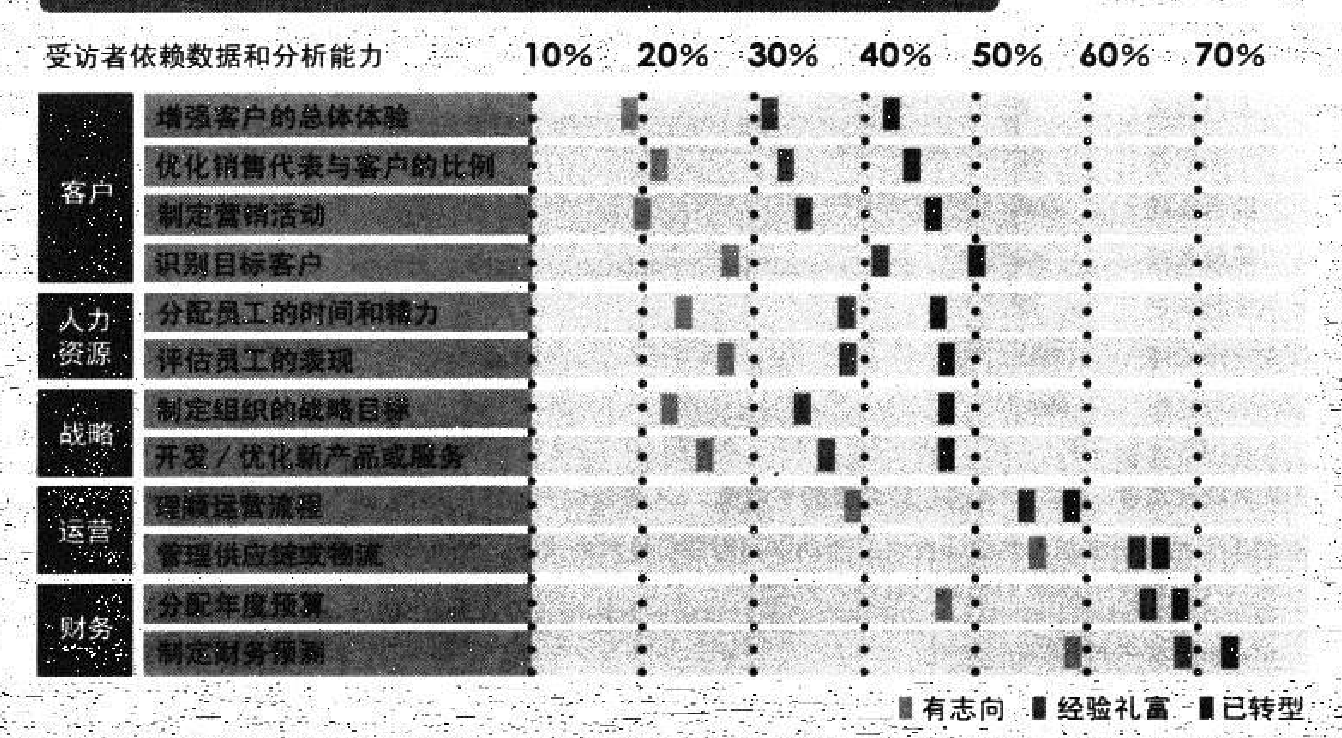
而表现较差的企业则是在客户服务、产品研发、风险管理、品牌管理、生产力计划等大多数的项目都倾向利用直觉进行判断。这项调查也发现，在“利用分析技术指导未来策略”上，表现优秀者（45%）比表现



较差者（20%）高出一倍，在“利用分析技术指导日常运作”的比例上，表现优秀者（53%）比表现较差者（27%）也高出一倍。

到了 2011 年，这项调查将样本数扩大到 4,500 份，调查发现有 58% 的企业运用分析技术创造竞争优势（2010 年的比例为 37%），而且应用程度较深的“已转型”企业和“经验丰富”的企业，和仅是“有意愿转型”的企业之间，绩效差距持续扩大。同时，大多数企业较依赖分析技术做出关于财务和运营活动的决策，但已转型的企业则是在客户面、人力资源面和策略面，使用分析技术的比例较高（见图 3-5）。

图 3-5　企业依赖数据资料和分析技术的程度



注：企业依赖数据和分析技术执行上述活动的比例。选择范围从 1= 直觉 / 经验、3= 经验 / 数据各占一半、5= 数据 / 分析。

数据来源：MIT 史隆管理评论与 IBM 商业价值研究院分析研究合作项目（Massachusetts Institute of Techno-logy 2011）



管理顾问公司麦肯锡（McKinsey & Company）不只一次指出，利用数据分析来指导决策的企业，比起一般企业生产力更高，净资产收益率也更高，其中尤以“联网型组织”最具优势。联网型组织指的是开放内部的信息渠道以及通过网络数据交流，让客户和供货商共同参与的企业。

换言之，大数据分析将是企业的新资产，就如同强大的品牌一样，是形成竞争力的重要基础。你所拥有的数据越广大、越多元，你驾驭这些数据的能力越高，就越能在市场胜出。

在大数据分析越来越被重视的情况下，企业和组织急需新技能、新视野，以及新人才。麦肯锡全球研究院（McKinsey Global Institute，MGI）研究报告显示，单是美国就需要14万到19万名拥有深度分析专长的工作者，以及150万名精通数据的经理人。

这让我想到了一个有趣的例子，还记得18年前风靡一时的美国情景喜剧《六人行》吗？这部由美国NBC电视台制播，从1994年开播到2004年才落幕的经典之作，前几季的内容中一直都有个谜团，那就是钱德勒（Chandler Muriel Bing）到底是做什么的？

剧中，钱德勒好几次向朋友们解释他的工作内容，表明他的工作职称是an executive specializing in statistical analysis and data reconfigu-ration（中文译为“统计分析和数据重组行政专员”），但还是完全没有人知道他到底靠什么维生，甚至连职位都没人记得住。

这个桥段成为剧中的经典笑料，一次在好友“琐事比赛”中，瑞秋真的不记得钱德勒的工作，只好乱编一个新词，叫他“ranspondster”；莫



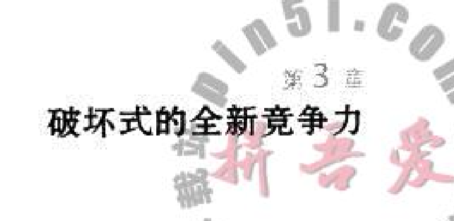
妮卡坦承好几次钱德勒在谈论他的工作时，她都在吹牛；乔伊也曾对着钱德勒大声斥责，“你竟然还敢叫自己会计师？”而当时钱德勒对此也只能疑惑地回答，“我的确不是啊！”

的确，在十几年前，多数人还不知道数据资料分析可以做些什么的时候，想要让别人理解这个工作的内容，的确不是一件容易的事情。“生不逢时”的钱德勒最终在剧中辞了工作，改当广告文案。但是到了现在，钱德勒的工作已成为硅谷最炙手可热的热门职缺。

美国《财富》杂志（Fortune）发现美国在超过 9%的高失业率的恶劣情况下，数据科学家（data scientist）却成为各个公司大力网罗的人才。数据科学家的主要工作内容，就是为企业分析每天搜集的大数据信息，从内部的销售报告，到客户发布的 Twitter 消息都包含在内。Google、亚马逊（Amazon）、沃尔玛等公司常年设置这个职位，而斯坦福大学（Stanford University）开办的数据采集课程也相当热门，去年有超过 120 名学生报名上这门课；不过 5 年前这门课程首次开放时，仅仅招揽到 20 名学生。

数据分析突然从索然无味的数学领域，变成抢手的主流学科，这代表的是，人们已经认知到数字和统计学是有趣的、实用的，甚至是很酷、很时尚的。此外，大数据分析也促使一种奠基于数据的新商业形态在市场上出现，我们称之为“数据经济”。

例如，一家运输公司在经营过程中必须收集全球产品运输的大量信息，于是创办了一个业务部门，专门为其他企业和经济预测单位提供辅助信息；电信公司通过数以千万计的客户数据，以及人们在社交媒体上



的活动，利用大数据分析发掘出多种使用者的行为和趋势，将这类数据和数据分析结果卖给有需要的企业。在越来越多企业或组织意识到数据的价值之后，这类贩卖数据、分析数据的产业势必将成为下一个热门行业。

小结

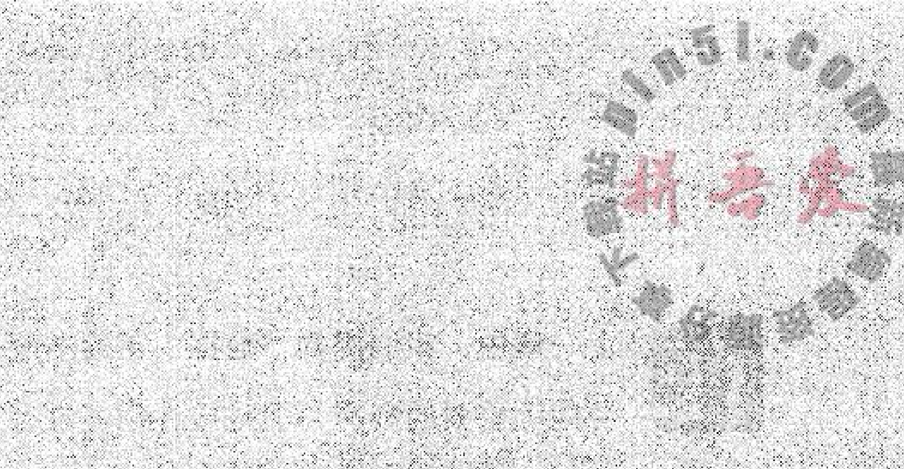
在大数据分析的浪潮下，世界的面貌即将改变，“数据”将登上舞台的中央。现阶段，我们已经可以从所有的终端设备获取数据，然后加以分析，最后做出实时的决策，例如从智慧水表、电表的传感器中提取数据流（data stream），利用这些实时数据节约40%的能源。那么未来，大数据分析还可以为我们做些什么？

根据麦肯锡估算，如果企业或组织能充分利用大数据分析，每年它可以为美国医疗业带来3000亿美元，为欧洲的公共部门带来2500亿欧元的潜在价值，零售业也可因此将其利润提高60%以上。因为对企业或组织来说，当你可以用实时或近乎实时的速度，整合来自不同渠道的庞杂数据，并运用强大的计算能力分析和挖掘，趋势就像是一幅图片，不再只是像素，而是一块拼图，让我们可以理解它；如果可以理解它，我们甚至就可以预测它的发生轨迹。

尽管目前大数据分析的市场仍在襁褓阶段，但它最终将会成为决定企业、组织，甚至国家（不仅仅是企业）竞争力的关键因素。更进一步思考，我们所做的每件事情、每个行动，都源自于我们自身的经验，而



这些经验都是数据资料，它帮助我们在生活中做出决定，所以就某种意义上来说，整个世界都需要做数据分析，不是吗？人类经历了铁器时代、原子时代和计算机时代，现在我们正在步入真正的“分析时代”；或许在未来，人们将会这么称呼它。



第4章 应用案例：从营销到反恐



在上一章中，我们提到“数据”正逐渐成为与实体资本和人力资源同等重要的资产，但是或许你还没有意识到，我们每一个人手边拥有的数据资料有多大的价值。

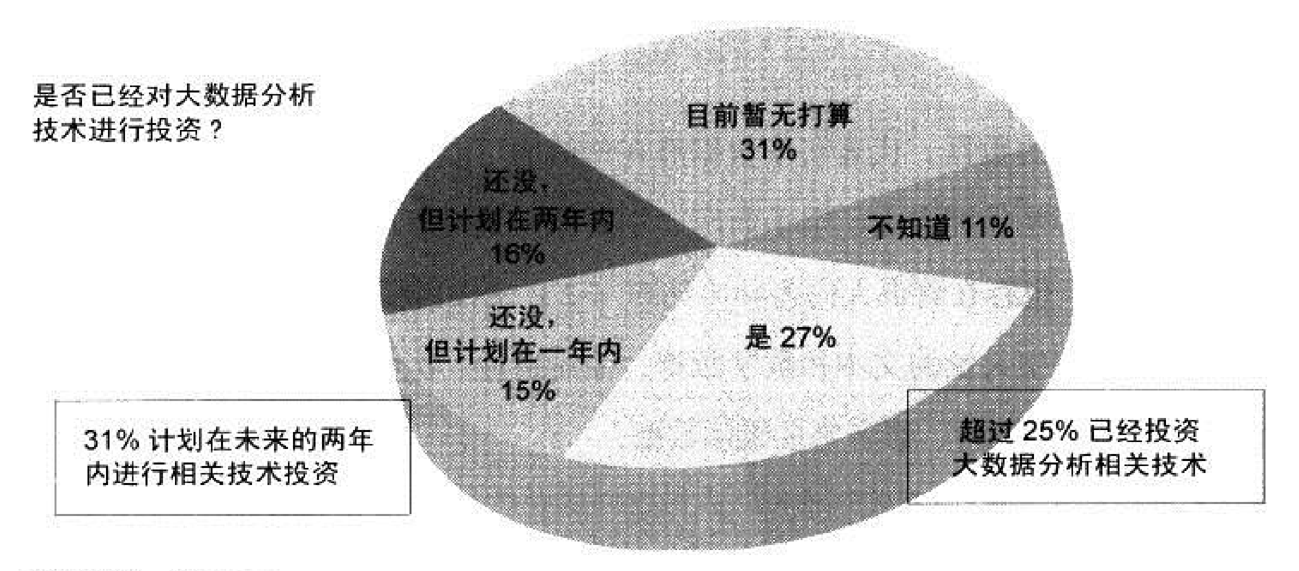
根据网络安全技术公司 McAfee 估算，一个网络用户拥有的数字资产（digital assets）平均价值为 37 438 美元。这些配置在计算机、平板电脑和智能手机里的数字资产，包括娱乐档案（例如音乐下载）、个人记忆（例如数码相片）、个人通信（例如电子邮件或日历）、个人记录（如卫生、金融、保险等数字数据）、职业信息（例如简历、投资组合、电子邮件联系人），以及你的兴趣爱好和创作。

以全球平均水平来看，一位受访者至少有 2777 份数字档案，其中与工作有关的数据资料价值最高，平均为 3798 美元，与个人嗜好有关的数据次之，价值约为 2848 美元，个人通信簿价值约为 2825 美元，娱乐档案价值约为 2092 美元。

为什么个人数字数据可以有这么高的价值？原因就在于它已被视为企业的战略性资产，企业可以运用这些个人数字数据开发出差异化的商品，借此改变用户的体验进而影响销售额。同时这些用户的使用或购买又可以帮助企业累积更多的个人数字数据，形成由数据驱动的经济循环。

因此，有越来越多企业愿意投资在大数据分析上，根据 Gartner 调查，目前有 27%的企业已经投资了大数据分析系统，也有 31%的企业预计在接下来的 2 年内投资（见图 4-1），希望能加强大数据分析相关硬件和软件的能力。

图 4-1 企业越来越愿意投资大数据分析



数据来源：Gartner

企业：预测使用者行为

不过，目前对大多数企业来说，所利用的数据资料量，还不到所获得的 5%。IBM 首席科学家强纳斯（Jeff Jonas）就预测，在大数据分析时代很多企业现在还可以弄懂 7%的企业数据，但这个数字很快就会下降到 4%，然后继续螺旋式下降。

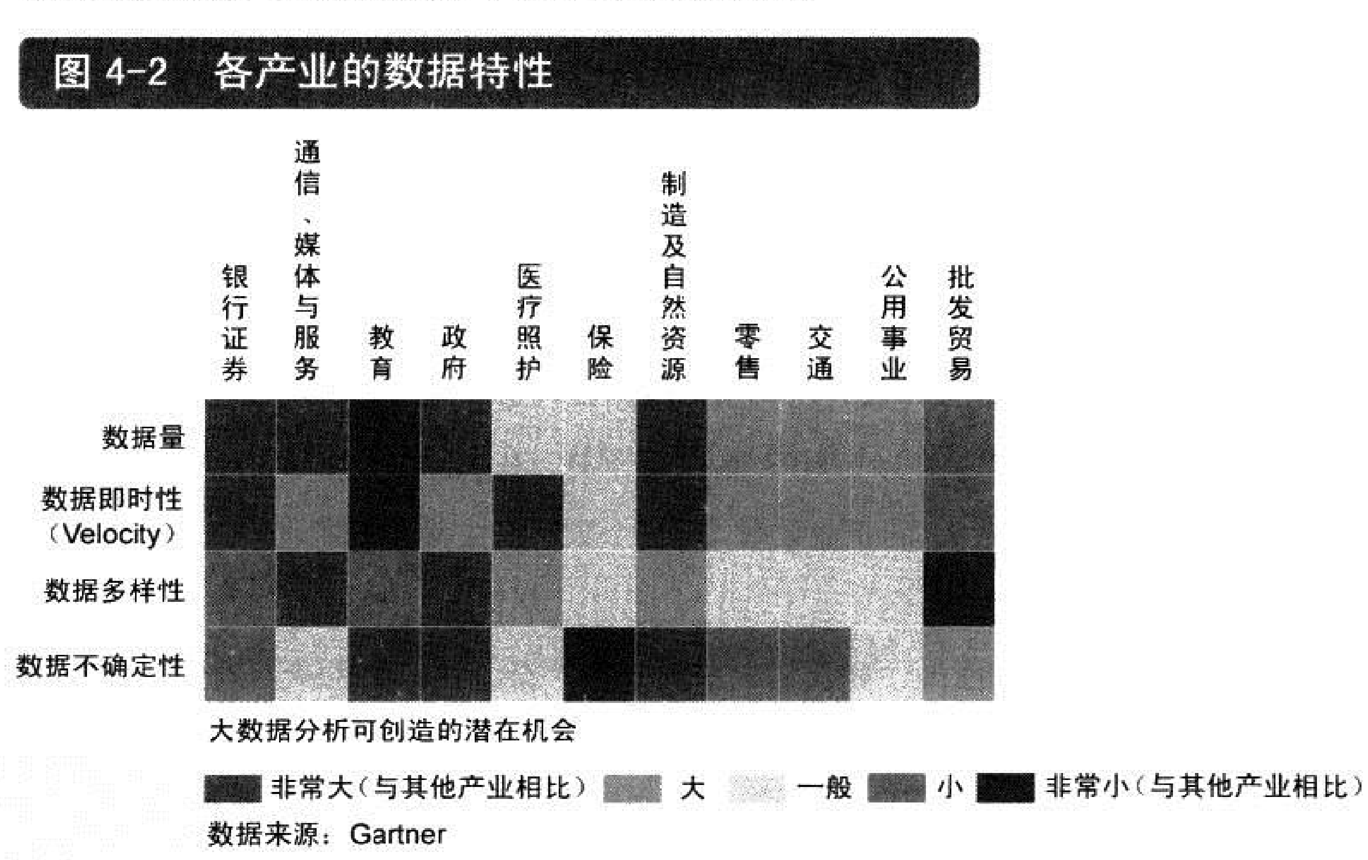
数据价值尚未完全被挖掘，和我们之前提到大、杂、快、疑的大数据出现有关，这也是企业为什么需要大数据分析的第一个原因。从数据量来看，各个产业拥有的数据资源（data resoures）其实并不少，根据麦肯锡全球研究院统计，光是 2009 年美国各行业储存的数据资料，



拥有 1000 名员工以上的单家公司就至少存有 150 太字节（TB）的数据，而产业总储存量更高达 6888 拍字节（PB）。有些产业则拥有相较于其他公司更多的数据资料，例如金融、媒体、医疗产业，以及政府部门。

照理说，拥有更多的数据，代表获取数据潜在价值的可能性越大。但是实际上企业常常却是坐拥宝山、空手而回，主要原因是在不同产业中所产生的数据类型，也存在着很大的差异（见图 4-2），例如金融业、政府部门、零售业会产生大量的文本和数字数据，而制造业、医疗业、媒体业则是会产生大量的图片、影音等多媒体数据。

图 4-2 各产业的数据特性



数据来源：Gartner

多媒体数据的增生，让企业即使储放了好几“吨”的数据，却可

能无法用“年”来排序、检查这些数据，更遑论分析了。这种情况使得企业开始思考数据的整合和应用，希望能借由大数据分析系统从这些“新品种”的数据和内容中，获取有意义的信息，并且创造新的机会。

第 2 个企业开始关注大数据分析的原因是，市场权力已逐渐转移到消费者的身上，企业希望借由大数据分析更加了解它的客户。就如同索尼（Sony）创办人出井伸之解释索尼衰落的原因时所说，“基于网络时代DNA 而生成的新一代企业，其核心能力在于利用新模式和新技术更加贴近消费者，并且深刻了解他们的需求，借此有效分析信息并做出预判，而未来传统的制造公司都只能沦为这种新型用户平台公司的附庸，这样的趋势不是靠管理就能够扭转的。”

过去，营销人员都将精力集中在大众市场；从制造业的角度来看，事情非常简单。以美国为例，凯迪拉克属于有钱人，中产阶级喜欢雪佛兰，庞狄克则是为喜欢耍酷炫耀的年轻人而设计的，而农夫，则买小货车就够用了。所以，我们回顾 20 世纪的制造业，其最高准则就是“大量生产”，不管是你吃的巧克力、穿的牛仔裤或是开的小轿车，都是工厂大量制造的产品，然后借由大众媒体让一般人了解这些产品。

这种求取最大经济规模的营销模式，在二次世界大战之后的数十年间，从美国开始席卷了欧洲、亚洲和全世界大部分的区域。在这样的工业时代里，广告宣传只要把所有人根据收入、性别和居住地区，区分为 5 类或 6 类，然后砸钱在他们阅读的杂志或观赏的电视节目上打广告，因



为产品几乎都是大同小异，谁的知名度高、谁的品牌大，谁就获得消费者的青睐。

但现在，这一切已经改变了。计算机的普及让制造流程拥有更大的弹性，只要一个简单的指令，饮料生产线就可以马上把柠檬汽水变成橘子汽水，织布机可以从条纹图案改为格子图案。由于制造商可以迅速把同一种产品做出数十种口味、形态的变化，让消费者开始有除了“价格”以外的选择权。今天的消费者要的是以适当的价格，提供我们想要的口味、颜色、质地或触感。

消费市场正从企业主导的经济型态，转变为由消费者驱动的经济型态，这表示营销人员必须把每一个人当成不同的营销目标。为了了解这种广大，却又单一的新消费观，营销人员不可能挨家挨户登门拜访；但是他们可以借由追踪和分析各种渠道中不断流出的数据，更贴近、理解和预测使用者的行为和需求，就像是广告业的名言：“你知道有 50%的投入被浪费了，只是你不知道是哪 50%。”而现在，企业可以借由大数据分析清楚看到那 50%在哪里，甚至掌握顾客消费行为，挑起消费者下一次购买欲。

政府：最具潜力的应用领域

当然，大数据中的潜在价值，绝非只能应用在商业上。麦肯锡全球研究院比较了各个产业可从大数据中获得的潜在价值，发现以批发贸易业、信息产业、金融保险业和政府部门的总体获益最大（见图 4-3）。

图 4-3　各产业运用大数据所得到的潜在价值

产业	部门	总体潜在价值指数	每家企业数据量	表现变异性	客户与供应商密集度	数据处理密集度	变动
商品	制造业						
	建筑业						
	天然资源						
	电脑与电子产品						
	房地产、租赁						
	批发贸易						
	信息产业						
服务	运输和仓库						
	零售贸易						
	行政、垃圾处理、整治服务						
	住宿和餐饮服务						
	其他服务（公共行政除外）						
	艺术、娱乐、休闲						
	金融保险						
	专业、科技、技术性服务						
	公司企业管理						
受管制/公众事业	政府						
	教育						
	医疗保健、社会救济						
	公用事业						

前 20%（潜力最大）
20% ～ 40%
40% ～ 60%
60% ～ 80%
最后 20%（潜力最小）
无数据

资料来源：麦肯锡全球研究院



信息产业名列其中，应该丝毫不让人觉得奇怪，因为这个产业本身就是数据资料密集，且分析技术较为创新和领先；而金融保险业则是本身数据资料的质量较好，拥有许多已经处理过的结构化数据。至于，包括批发贸易、医疗业、制造业和零售业，他们零碎的产业结构特性，最能阐释大数据大、杂、快、疑的特性，但是如果小公司可以创造、获取、利用大型数据池，来分析数据并得到洞察，反而更容易创造价值。

最有开发潜力的一块则是政府的公共部门。政府收集了数千万、数亿笔的民众数据，而且可信度比一般商业交易来得高。但是公共部门除了储存以外，很少管理和利用这个宝贵的数据资源。美国政府是将大数据分析用于公共安全管理的先行者，最经典的案例就是美国最大的警察局——纽约市警局（NYPD）。

纽约，一个被称为集天堂和地狱于一身的城市，它是全世界的金融和商业中心，也是美国人口数量最多、密度最大、多元化程度最高的城市，但它也是著名的犯罪之都。在 20 世纪 90 年代，在纽约平均每天会有 6 个人死于犯罪，每小时有 16 辆车子不翼而飞。

身为这个城市的守护者，纽约市警局的正式警官超过 3.7 万人，辖下包括多个特殊职能部门，利如战术行动小队、港口巡逻、空中支持、拆弹小组、反恐小组、罪犯情报、打黑、缉毒等，另外还有专门协助打击计算机犯罪的小组。他们拥有全美最丰富的犯罪数据库，也是全世界第一个运用数据分析系统打击犯罪和预测犯罪的单位。

这个单位的产生源于 20 世纪 70 年代，当时为了研究地铁抢劫案的



发生规律，纽约市警局开始以地图和不同颜色的大头针，针对地铁抢案进行人工分析，预测可能发生抢劫的时间和地点，这个方法成功让地铁抢案发生率下降了27%。

之后，这个当时被称为“未来图表”（Charts of the Future）的统计法，开始大量运用在偷盗、毒品交易、帮派械斗等不同的案件当中，而以“改善治安”为主诉求的前市长朱利安尼，1994年上任后立刻指示纽约市警局开发一套电子版的“未来图表”系统，现在这套以地图为基础的统计分析系统“CompStat”（Computer Statistics，计算机统计），在警界已经演变成为一个专有名词。

在网络尚未普及的20世纪90年代，工作人员每天通过电话和传真向全纽约76个警区收集数据资料，再将数据统一编录到CompStat系统进行分析。然后，每周两次招集全警区的指挥官，发布最新案件在各个辖区的地图上的位置和代表意义，进行未来应对对策以及警力调配的模拟。次年，凶杀案下降了25%，车辆窃盗案下降了24%，也让朱利安尼的人气大为攀升。

2001年震惊世界的9.11恐怖攻击事件发生后，美国政府开始以数据资料分析为基础，在CIA、FBI以及各大城市警局等公共安全管理单位，建立新的反恐和保安系统。国防部和IBM合作研发了“流计算”系统，可以针对动态性质的大数据，略过数据仓库处理，直接加载多元结构分析，进行实时性、高复杂度的分析，在微秒之内生成分析结果，将数据分析系统在公共安全的运用上往前推进了一大步。

由于在9.11事件中发现，19个劫机者当中至少有11个人持假身份



证件入境，因此这套系统最早是用于个人出入海关时的脸部识别，后来也被运用在街头巡逻警车上。警察在巡逻时扫描经过的车辆，就能立即知道这辆车是否被列入失窃名单，若发现可疑人士，系统会将车辆数据比对其他不同来源的数据，以预测犯罪活动。

例如，一个初入境的外国男子，买了一张从开罗到洛杉矶的单程机票，在洛城当地租了一辆车，并在当地一家俄罗斯银行提取了一大笔钱，其间还频繁地和叙利亚方面电话联系，然后独自一人去了迪斯尼乐园。原本这些蛛丝马迹并没有被串连起来，后来因为他前往迪斯尼的路上超速，被停在路旁的交警把车辆数据记录了下来，并对该名男子盘问了一番。

盘问过程中，交警觉得这个人相当可疑，于是把数据回传总部，系统开始比对出入境数据、提款记录以及电话记录，发现他的资金往来和联系对象的确非常可疑；进一步比对 CIA 开罗办事处收集到的指纹和 DNA 样本，以及他在俄罗斯、开罗、叙利亚等国外账户的交易情形，加上调阅了迪斯尼乐园的监视器画面，发现该名男子到游乐园根本没有玩乐，而是花大量的时间拍照和记录；于是警方联合了 CIA 采取行动，提前阻止了一场即将发生的恐怖袭击。

这是 CIA 所提供的数据，故事中的主角和地点当然经过了修改，但这也代表数据分析系统的功能，对于公共安全管理的贡献。目前包括 CIA、FBI 和各国警方都开始利用类似的数据分析系统，预测未来的恐怖袭击和重大犯罪，并成功破获了像是贩毒组织和非法交易儿童等重大案件。

除了公共安全管理之外，公共行政效率是数据分析系统应用的另一大重点。美国政府一开始实行是为了弊端丛生的医疗保健制度。美国现行规模最大的两项医疗福利计划，是 1965 年通过的医疗保险计划（Medicare）和医疗补助计划（Medicaid）。前者专门针对残障人士和 65 岁以上的老人，费用通过保险来支付，由联邦医疗保险和补助中心（CMS）监管实施；后者则是以贫困人口为对象，费用由政府直接支付，由 CMS 中心和各个州政府共同实施。

然而，这两项福利政策最大的争议在于，是否会因为浮报滥用而导致全民供养装穷的懒人，造成一般民众看病就诊必须支付高额的费用。因为根据 2008 年联邦调查局在《财务犯罪年度公开报告》中估计，政府每年的医疗开支中有 3%～10%涉嫌造假和欺诈，而上述两项医疗福利政策涉及 1 亿人，其中的虚报账单、重复申报、隐瞒收入和存款等事件层出不穷。

CMS 决定通过数据资料来抓弊，2001 年由加州率先推出“保险补助双向核对”（Medical-Medicaid Data Match）制度，整合两项福利政策的数据，比对其中的人员、时间、价格、地点等，以分析系统自动确认矛盾、异于常态的支付记录，一旦发现造假或者申报不实就开始进行人工追讨。2004 年开始在各州扩大实施，施行一年多之后至少追讨回 1500 万美元的超额申报。之后除了 CMS 中心，联邦政府的各项社会福利项目都开始陆续采用这种做法，预计 10 年内可以为国家节省 2000 亿美元。

除了节省经费以外，各国政府或组织也开始尝试利用大数据分析协



助救援行动和预测经济。联合国在2010年的海地大地震中，以追踪海地人所持手机内部的SIM卡信号，分析出逾60万名海地人逃离太子港之后的目的地，提高了救援的效率。后来，当海地爆发霍乱疫情时，同一批研究人员再次通过SIM卡追踪，把药品投放到正确的地点，也成功阻止了疫情的蔓延。

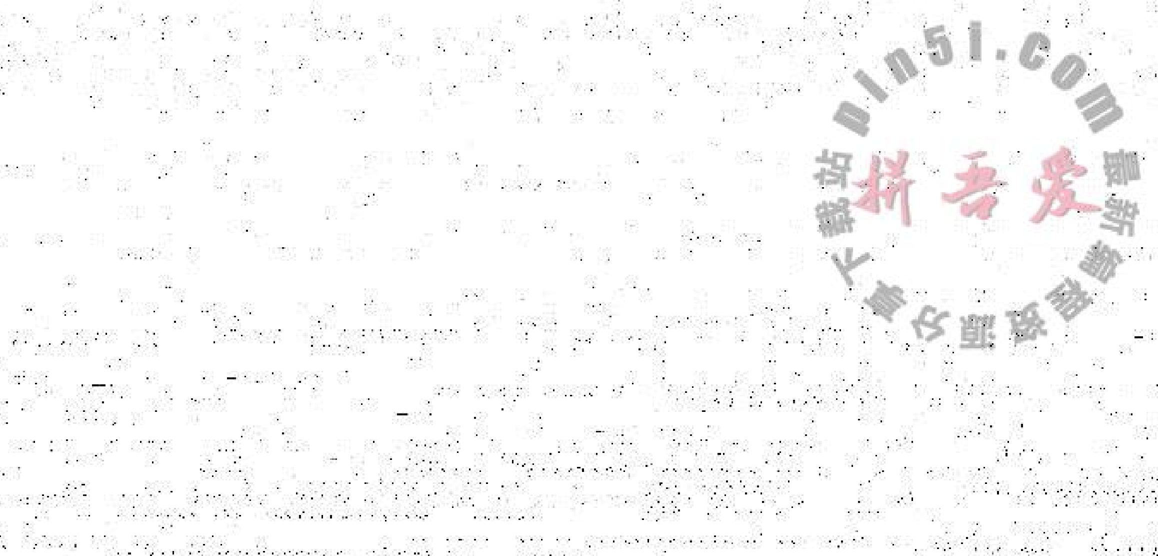
此外，我们在第1章提过的联合国“全球脉动”（UN Global Pulse）研究计划，也借由社交媒体网站每天产生的2.5艾字节（EB）数据，考察公共议题的风向、识别经济趋势并预测即将显现的问题。目前他们已经利用“面包实时在线价格”（Real-Time E-pricing of Bread），在6个拉丁美洲国家建立每日价格指数，提早测知通货膨胀的变化。

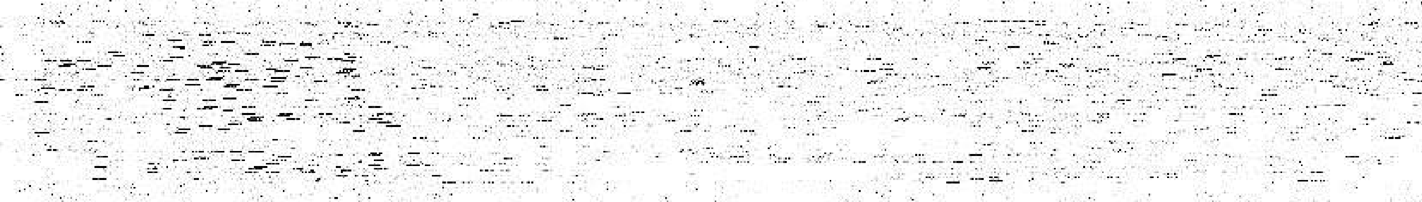
当前，我们的世界正变得越来越不稳定，过去几年的全球经济体系也正变得越来越动荡。不管是政府或企业决策者都已经意识到这些不断发生的危机所带来的昂贵代价。他们想要判断当危机发生时，在哪些地方、什么时间、哪些群体会受到多严重的影响，并且知道如何预防或是怎么将损失控制在最小范围内，也开始注意到“大数据驱动政策”的可行性。

尤其是从公元前6世纪的亚历山大图书馆，到无所不搜的Google，再到即将成形的SoLoMo趋势[1]，现在我们可以取得的数据内容已经变得

[1] SoLoMo是2011年2月，著名的KPCB风险投资公司（KPCB）约翰·杜尔（John Doerr）所提出的网络未来趋势。So=Social，指的是由Facebook、Twitter、Zynga这些公司发起的社交化运动，而未来网络应用必须要靠着社交网站来提供更多价值。Lo=Local，指的是随着智能手机的普及，我们得到的数据将会越来越本地化、因地制宜，因此要开发的是基于位置的服务（LBS，Location-based Service）。Mo=Mobile，则是指移动互联网（Mobile Internet）的崛起，预计会在2013年超越台式计算机，成为人们上网的主流方式。

越来越细致，也越来越个性化。如今，我们可以利用极大、极丰富的数据资源（包括旧数据和新数据），针对社会人口、市场趋势进行前所未有的实时、个性化分析，这不仅仅是成为企业提升核心竞争力的有效方式，更提供了一种保护人类社会、协助全球发展的公共能力，接下来将针对各个产业和政府的公共部门，让你了解他们怎么利用大数据，让这个世界更安全、更美好、更具智慧。







第5章 零售：更好、更快、更便宜



回顾历史轨迹，零售业和经济增长息息相关。然而近 10 年来，尤其是在金融海啸过后，整体零售业的增长已经趋缓。根据 MGI 调查，人们花钱的额度正缓慢地减少。以 1 美元来说，1999 年人们花在零售业的钱约有 0.6 美元，但是到了 2008 年却只有 0.31 美元；这不仅显示了人们对于日常用品的“价格”开始精打细算，也代表零售业获取利润的压力日渐升高。

同时，随着各国的发达程度增高，就连之前被视为世界工厂的中国，也逐渐转而成为世界市场。低人力成本的优势不再，制造业的利润几乎被榨到极致，为了减少不必要的进货成本，零售商转而从消费者需求着手，希望更精准地在市场上提供对的产品。

不过，因为电子商务和行动商务崛起，2012 年以后将有 50%的零售业会跨入在线购物或是被在线购物所影响。另一方面，现在消费者选购产品时的参考依据，网络信息已经取代了传统的广告。在网络易于比较商品信息和价格的特性下，零售业必须想办法让产品贴近消费者的需求。但是随着消费形式越来越复杂，零售商可能必须同时获得直销、编目、门市和在线购物等不同来源的数据资料，并了解其交互作用的脉络和意义，才能得知消费需求的全貌。

在消费者相关数据源量大又纷杂、且难以整合的情况下，在营销、采购和供应链管理上运用大数据分析，“整合分散的数据源”、“结合内外部数据”及“网络内容分析”将最具潜在效益（见图 5-1）。根据 MGI 估计，大数据分析每年将可提升零售业 0.5%的生产率；而成为大数据分析先锋者的单一公司，借此增加生产力、降低成本及改善营运效率之后，将可以增加超过 60%的营利率。

图 5-1　零售业采用大数据分析的潜在效益

大数据分析在零售业不同的部门有不同的潜在价值

		保险、护理	一般商品	建筑材料和园艺	非实体店面	餐饮	服饰配件	运动、兴趣、书籍、音乐	3C家电	家具和居家装潢	其他
市场营销	增加交叉销售										
	定点营销										
	店内行为分析										
	客户微区隔										
	情绪分析										
	改善多渠道体验										
商品推销	商品组合优化										
	定价优化										
	商品摆放、设计优化										
营运	表现透明度										
	劳力投入优化										
供应链	存货管理改进										
	物流优化										
	告知供应商协商内容										
新企业	价格比较服务										

■ 适用性最高　░ 适用性最低　注：网络市场难以量化，因此并没有包含在此。

数据来源：麦肯锡全球研究院



整合分散的数据源

位于美国俄亥俄州的辛辛那提动物园，是美国历史最悠久的动物园之一（1875 年开业），占地约 40.5 万平方米的园区里包括超过 500 种动物和 3000 种植物。它同时也是美国第一家采用无隔离方式和游客互动的动物园，每年平均游客量超过 120 万人次。

2009 年，辛辛那提动物园遭遇了前所未有的经营危机。当时正值金融海啸发生后不久，全美陷入经济大衰退，动物园、水族馆、博物馆可以拿到的政府补助因为税收减少而大幅减少，而且到动物园的游客量以及游客们游园时的花费也大幅降低。另一方面，动物园会员人数不但减少了，入园后的花费也远不如非会员的游客。

对于一个营运费用有 50%要依靠税收的非营利组织来说，每年 2600 万美元的年度预算中，必须自筹资金达 1600 万美元以上。他们要如何自力更生，又不给当地纳税人造成过多负担？营运团队决定借由商业分析（Business Analytics）减少开支，并增加入园费、会员人数和食物、园区商品等现场销售的金额。但是，问题在于进行分析所需要的数据！

以往，动物园的营运数据来自于入口处、贩卖亭、游客中心等 3 个不同的来源，而且都是由员工手抄记录，然后一个星期汇整一次。所以，即使累积了大量的数据，也无法进行整体分析，并帮助现场人员进行实时性的资源调配。占地广大成了另一项难题，由于园内设置了多个销售点，管理阶层要走遍整个约 40.5 万平方米大小的动物园，才能知道会员

们在浏览景点时做了什么，游客是如何花钱的，因此常常不知道应该提供什么样的产品和服务。

经过调查比较，辛辛那提动物园最后选择和 IBM 合作，设计了一套横跨票务、零售和食品贩卖的数据整合分析平台，汇整以往的历史数据，同时在现场服务的每位员工也都必须把每一个入园游客，和每一笔商品事务数据编码记录，实时回传到这个平台。

经过数据分析之后，园方发现以往动物园给当地居民很高的入园折扣，但其实这些会员却不常利用，因此决议删除这项宣传费用，改以会员购买园内商品时给予折扣，果然因此提升了商品销售业绩。另一方面，园方也减少了年度广告预算，改以利用经过编码记录后得到的游客数据，寄送宣传品给邻近地区到动物园频率较高、收入较高的游客，此举不但降低了 43%的广告费用，而且提升了 4.2%的整体售票率。

另外，以往贩卖亭的进货方式，都是员工以目测的方式决定要不要进货，或是进哪些货；经过数据分析之后，园方发现有些食物会在特定时段卖得特别好。例如在即将关园的这一段时间内，就是冰淇淋的销售高峰期，但却常因为存货不足而引起顾客抱怨，而某些相关商品滞销已久却没人处理。于是他们依照销售热潮分批进行补货，剔除滞销商品，并提高热销商品的价格，也让食品和商品销售业绩双双增长了 18%。

这一套数据整合分析平台，也辅以 iPad 和智能手机打造一个移动商业分析的架构，管理阶层不用走遍园区，就可以将游客数据和园内销售数据集中到 iPad 上的企业仪表板（Dashboard），并据此调整营销模式和资源分配，系统也可以主动利用智能手机让每个员工快速得到重要信息，



当会员入园时也会接到通知，让他们知道今日活动或促销优惠。

实行后第一年，辛辛那提动物园在不景气中增加了 35 万美元的收入、5 万人次的新游客，现在因为这一套系统得到的平均年收益为 73.8 万美元，投资回报率（ROI）高达 411%，投资回收期仅 3 个月，而且没有增加任何一名员工和人事成本，甚至员工们因为不再需要走动式人工记录数据，可以有更多的时间管理贩卖服务、会员服务，以及资金募集的事务。

结合内外部数据

Sun World International 是美国一家专门种植和运输新鲜农产品的中型企业，它从蔬果包装业务开始，于 20 世纪 80 年代跨足生产，目前拥有超过 6400 公顷的耕地，并和 950 位签约农民合作，主要生产作物包括甜椒、葡萄、核果和蔬菜，每年出货量高达 1100 万箱，除了美国本地之外，还外销到意大利、澳大利亚、智利、墨西哥和南非等地。

食品安全和创新育种让 Sun World 闻名业界。例如，即使到了 9 月，你还可以在占地 73 公顷的实验农场里看得到无籽红葡萄，就是该公司借由品种改良延长水果生长季的成果。然而，因为全球化商业的影响，水果和蔬菜等农产品都能在短时间内，通过空运送达地球的另一端，再加上农产品容易因为气候影响而导致某种作物在同一时间内大量采收，以致于价格暴跌，因此以往 Sun World 4 个种植区的农场经理每个月会到总公司开会，研讨之后的预算、成本和生产率。

总公司利用 ERP（企业资源规划）系统提供不同产品类型、品种和地区的销售、成本及利润数据，但在当前全球化市场风险增加的情况下，Sun World 需要降低成本并提高生产率，以便在国际采购市场上更具价格优势，所以除了既有的公司 ERP 系统之外，Sun World 希望能够纳入像是水成本、燃料成本和消费者购买模式变化等参数，将内外部的数据结合后进行分析，做为农产品种植结构的依据。

Sun World 和 IBM 合作，进一步借助数据分析系统重新设计生产流程。但是一开始，比起数据，经验丰富的农民更忠于直觉，他们认为自己比起那些冷冰冰的数据更了解在哪些地区、哪个时节实施灌溉和施肥效果最好。因此系统导入初期仅用于灌溉系统，以基于降低用水成本和促进农作物营养的数据分析结果，在水资源短缺的区域以高效滴灌法替代传统的地面灌溉。没想到，效益出乎农民们意料的高，以葡萄来说，收割成本降低了 5%，并减少 20%的燃料使用量，同时产量更是大幅增加 50%。

如今，农场经理们已经非常习惯使用他们的手机或 PDA，在收成期随时察看每一个员工或农场的生产率指针，而且因为数据分析，现在 Sun World 农场的人力成本减少了 550 万美元，每一亩耕地的利润增加了 8 倍。此外，为了提高销售人员和客户协商时的谈判优势，这套系统也设计了实时信息显示屏幕（ticker display），可以不断更新全世界最新的农产品价格，也让 Sun World 在波动的零售市场中，客户数量每年增加 20%。

另一家利用内外部数据提高效率的中型企业是美国老字号的 Papa



Gino's 比萨连锁店。这家强调正宗意大利家庭食谱的比萨店，从 1961 年在波士顿开设的家庭餐馆，至今已成为在新英格兰地区拥有将近 200 家店的连锁企业，提供全系列的潜艇堡、沙拉、比萨和意大利面。1997 年 Papa Gino's 并购了百胜餐饮集团旗下的 D'Angelo 烤三明治店，目前展店数也已超过 100 家。

在淘汰率很高的餐饮业屹立超过 50 年，Papa Gino's 至今仍保留可供顾客观看比萨制作过程的透明厨房等经典元素，同时也不断求新求变，例如推出以 1 美元换取一个积分，50 个积分可以兑换 5 美元的忠诚卡。

不过，就像大多数餐饮业一样，Papa Gino's 的营运数据分散在 3 个不同的系统中，包括 ERP、POS 和 Excel，因此光是编制年度预算，从不同的电子表格中取得数据资料并整合到完成，就需要 4 个月。

随着店数越来越多，Papa Gino's 采用 IBM 的商业分析系统，建立一个可以整合所有数据的单一平台，来加快他们的营运效率。除了新的财报标准和预算编制过程可以自动产生表单，让该公司财务人员增加了 25%的工作效率以外，这套系统也可以经过比较目前和前几年同期的销售数字，提供管理人员和区域经理两周后的人力需求预测，以确保顾客可以得到最好的服务。

同时，系统分析订单相关数据后也发现，原本 Papa Gino's 将外送交货时间订为 45 分钟，但实际上工作人员可以在不到 30 分钟的时间内交付；总公司依此调整交货时间，让客户的期望获得更大的满足。另外，系统还发现，比较网络、电话及其他订购渠道，网络顾客根据季节变化，

占了总订单的 40%～70%，因此 Papa Gino’s 也决定加强在线服务。

在营销活动上，以往 Papa Gino’s 的做法是以电子邮件、短信和广告邮件发送优惠券，以便对广大的消费者进行促销。数据分析系统追踪了这些优惠券实际使用的情况，并发现网络优惠活动可以促使顾客的用餐率提高 35%，在线销售额增长 50%，让从实体店面经营起的 Papa Gino’s，更加体会到网络促动消费的威力。目前他们除了推出 iPhone、黑莓机和 Android 移动平台的 APP 之外，也将数据分析扩及到 Facebook、Twitter 上，不久后还会纳入 Foursquare（链接社交和地理位置的网站）、Yelp（美国最大的点评网站）、Groupon（全球最大团购网站）和其他十几个社交网站的数据。

Papa Gino’s 相信，社交媒体和移动应用程序（APP），在餐饮业的未来发展上将有极大的影响力，而结合这些外部的网络内容数据，将会让这家年过半百的连锁餐饮企业更有活力和创造力，持续居于业界的领先地位。

分析网络内容，贴近消费者的心

在第一线直接面对广大的个体消费者，让零售业相较于其他行业，更希望能测知每一个消费者心目中的偏好，也因此对于大数据分析在网络内容上的应用上相当关注。对于零售商来说，由于网络上无法完全凭着商业意图控制消费者意见，所以成为目前最客观的消费者意见反映渠道，但是要如何让分散在社交网站帖子、各大讨论区的杂乱消息，可以



有意义的被利用？

全球第二大的卡夫食品澳洲分公司，就为了澳洲特有饮食文化中最著名的象征维吉酱（Vegemite），分析超过 10 亿条社交网站帖子，以及将近 50 万条论坛的讨论内容，为这种“国民食品”找到了新的市场价值。

维吉酱是一种黑褐色的抹酱，很粘稠但不拉丝，看起来有点像巧克力酱，但它的口味却是咸的！有人形容它和日本纳豆或是中国的臭豆腐一样，都是外地人难以理解、但当地民众却相当喜爱的食物。这种利用酿酒酵母开发的食品，从 1923 年开始就在澳洲和新西兰开始流行，而现在在澳洲，有 70%的人早上起床的第一件事，除了冲杯咖啡之外，就是在烤面包上抹上一层厚厚的维吉酱。

以一个成功的商品来说，维吉酱的味道和外观都是澳洲特有的，要改变其中任何一个特质都是不明智的决定。不过，卡夫食品希望为这个长青商品寻求其他消费机制，以延续这个产品的生命周期，于是他们决定为维吉酱推出一个新的大型宣传活动，叫做“你有多喜欢维吉酱？”。

以往，要在占地广大、种族多元的澳洲进行一项消费者营销调查，通常需要 4～6 个月才能完成，而且还无法囊括所有不同族族和文化背景的澳洲人。要通过这么狭隘的信息来改变一个已经成功的品牌，其实是非常冒险的，但是如果针对数以亿计的网络帖子，可能需要好几年才能收集完成。为了找寻一个可以收集各种背景、各种社经地位的消费者信息和意见的机制，然后将这些数据资料运用到有意义的方向来促进品牌

增长，卡夫食品采用了 IBM 的 COBRA 分析系统（Corporate Brand and Reputation Analysis，企业品牌和声誉分析），利用这种基于文本数据的先进分析工具，抓取了 10.5 亿条社交媒体帖子，以及 47.9 万条在论坛和讨论版内容中有关于维吉酱的信息，再针对这些非结构化数据进行深层原因分析，而且时间只花了短短 2 个月。

分析结果大大出乎卡夫的意料，在这些庞大的帖子讨论中，大约有 50 万人，用 38 种语言提到了维吉酱，而且大家谈论的焦点并不是维吉酱的咸度或是产品包装，而是除了抹在烤面包上以外各种各样不同的吃法，以及在国外怎么买到维吉酱！而且有数十万澳洲人以创新或不同的方式食用他们的维吉酱，例如抹在西红柿、鳄梨或奶酪上，最后分析出 32 种维吉酱新吃法。

同时，语义分析也显示，网络上发言的消费者绝大部分对维吉酱充满了感情，对消费者来说它不仅仅是一种食品，更是澳洲民族的一种象征。此外，语义分析也显示出消费者普遍关心的 3 个趋势：健康、素食主义和食品安全。这些分析结果对卡夫调整维吉酱的营销策略有很大的启发，例如在关于健康的讨论议题中，"叶酸"这个营养素的名称就被频繁提及。叶酸为人体代谢所需，对孕妇尤其重要，而这项发现也为卡夫食品未来打开孕妇消费者市场提供了依据。

现在，维吉酱的官方网站上已提供了许多不同的吃法，并持续邀请网友们进行调查，了解他们食用维吉酱的方式。论坛中还设有儿童专区，供孩子们参与讨论维吉酱的新吃法，以培育下一代消费者。这项成功的宣传活动让维吉咸味酱的销售额大幅增加，每个月销售量高达数千吨，



创造了该产品的历史新高纪录。最令卡夫食品相关人员兴奋的是，他们终于成功让澳洲妈妈们一次购买 2 罐维吉酱了！

利用网络内容成功贴近消费者的还有罗森（LAWSON）连锁便利商店。罗森是日本第二大便利商店，在日本的规模仅次于 7-Eleven，店数超过一万家；而在中国的上海、重庆和大连等城市也拥有将近 400 家超商。除了店数众多之外，出售的货品也很多元，包括杂志、漫画、饮品、药物、零食及便当等。

由于业务范围广大，加上消费者对便利商店的忠诚度不高，以及各地区消费者的品味不同，罗森一直希望借由一个可以实时挖掘文本数据的工具，收集客户在社交网站上的意见以及其他在线评论，以协助总部进行产品开发和服务改进。之前，罗森使用的数据分析系统以“关键词”为搜索目标，但是无法以分析历史数据来识别新兴偏好，而且因为不能自动搜索，最后还是需要手动改变关键词和记录结果，很难将发帖过程累积成可用的数据。

在旧系统无法符合总公司的需求下，罗森转而以 IBM 的内容分析软件来自动收集、累积和分析不同社交媒体和网站讨论区的大量数据，并将 IT 管理架构的流程集成起来，建立一个从信息处理引擎到所有商店信息系统的单一控制面板。

这套系统使用了自然语言技术，可针对包含英文、日文、中文等 11 种语言的非结构化和结构化数据进行分析，并且可在同一时间内自动收集超过 1000 万个网络帖子，自动分析结果产生后，系统还会显示警报提要。例如，系统发现网络上客户抱怨巧克力奶油冻不甜的频率增加，就会自

动开始收集一系列有关于“巧克力奶油冻的味道”，或是“不喜欢这种巧克力”等上下文义相关的评论和帖子，并将结果显示在每一部终端计算机的企业仪表板上，以帮助总部迅速采取行动。

例如，罗森日前针对年轻女性推出一款低热量新餐盒，包含一碗米饭和丰富的鸡肉、鸡蛋、蔬菜，但是销售却不如预期。经过内容分析后发现，二三十岁的女性喜欢这个餐盒的味道，却不喜欢新产品的包装。该公司更新了包装后，果然产品销售也跟着上升。

此外，这套解决方案也可以借由各类食品的保鲜期，监测产品新鲜度和浪费的情况，让店经理可据此控制商品库存量和鲜食进货量。实行之后罗森 2011 上半年度营收比前一年同期增加了 0.5%，当年度合并营利也打破了该公司历年纪录，预计 2015 年公司税前营收，将因此较 2011 年大幅增长 60%。

从上述案例中，我们可以知道，现在的消费者期望的是：better products（有更好的产品来符合他们的需求）、less time（可以花更少的时间去寻找）、right price（价格合理的好产品）。而在你身处的产业中，该怎么利用大数据分析，为消费者提供更好的质量、更快的速度和更经济的价格呢？以其他零售企业可以得到的效益来看，这绝对值得你花时间好好思考。

内容分析软体的基础架构

- 层面分析（Facet analysis）：列出包括经过频率、相关性和名称等条件挑选的关键词。
- 时间序列分析（Time-series analysis）：获取特定数据在指定时间内发生率的变化。
- 趋势分析（Trend analysis）：自动侦测关键词在一个时间序列内频率的显著增加或减少。
- 偏差分析（Deviation analysis）：比较和自动检测关键词在同一层面、同一时间范围内，频率的显著增加或减少。
- 双层分析（Facet pair analysis）：自动侦测两个任意的层面关键词之间的关联性。
- 连接分析（Connection analysis）：将同一网络上、两个不同层面的关联性可视化。
- 企业仪表板（Dashboard）：在同一个屏幕上显示多个分析图表。





第6章

医疗：降低成本，促进医学研发



对于美国，抑或是许多医疗系统较为完善的国家或地区来说，逐年增加的医疗成本已经为疲弱的整体经济带来很大的负担。以美国为例，随着老龄人口增多，医疗护理的从业人员势必也要跟着增加，甚至药物和仪器设备的需求也会变多，因此根据 MGI 估计，未来 20 年间，医疗业成本将会以每年 5%的速度逐步上升。

应对成本增加，收费也将会越来越贵。目前美国民众每人每年在医疗护理上面花的钱大约是 2500 美元，总体金额高达约 7500 亿美元，贵到生不起病已成为民众最诟病的社会问题。如果情况无法改善，MGI 估计未来 10 年美国民众每年医疗花费的增长率，将超过 GDP 增长率大约 2%。

如何提高医疗业的效率，减轻逐年增加的成本压力，为民众提供合理的医疗质量？目前已有医疗机构开始借助新科技收集数据以改善营运。例如美国的退伍军人事务部（The Department of Veterans Affairs）就成功利用电子病历、医疗信息技术以及远程病人监测技术，使其在病人护理、药物处方和临床治疗上的效果都较私人机构为佳；英国健康暨临床医学研究院（National Institute for Health and Clinical Excellence）率先使用大型临床数据，探讨新药物和昂贵治疗方法的成本和治疗效率，并以此做为与药厂协商价格的主要依据；意大利药品管理局（The Italian Medicines Agency）也正仿效这个做法，收集并分析使用昂贵新药的临床数据，以评估新药物进入医疗市场的价格和条件。

然而，绝大多数的医疗机构，借由新科技收集数据和改善营运的

程度远落后于其他行业。医疗业所需要的 4 大数据源包括临床数据（Clinical data）、药物和医疗用品研发数据（Pharmaceutical and medical products R&D data）、付费者作业和成本数据（Payor activity and cost data）、患者行为和情绪数据（Patient behavior and sentiment data）；其中只有药物和医疗用品研发相关数据目前数字化程度最高，未来可望与患者行为和情绪数据，或是个人基因数据库结合，帮助研发出更好的新药。

至于临床数据的部分，美国至少还有 30%的病历、实验和手术报告等临床数据尚未电子化。因此，2009 年提出的美国复苏与再投资法案（American Recovery and Reinvestment Act）中，就提供了 2000 亿美金的奖励基金，希望在 5 年内帮助所有诊所和医院全面将病历和临床数据数字化，以利未来的数据收集和分析应用。另外，由于付费者作业和成本数据分散在政府、保险公司、医疗院所、药商等各个单位的手中，不仅没有统一的标准格式，而且因相关数据在企业端大多被视为商业机密，所以也无法集结。

不过，医疗业若能增加 IT 投资和数据收集的能力，做好保护隐私的机制，并整合政府、医疗院所和企业的数据源，长期来看，MGI 估计这样做可为整体医疗业带来每年超过 3000 亿美元的新价值，减少大约 8%的国家医疗保健支出（以 2010 年为基准），并且增加 0.7%的生产力（见图 6-1）。而目前医疗业在数据分析的实际应用上，已经运用了实时监测并分析大量的仪器数据，帮助医疗机构和学术单位提高“临床试验”和“医学研发”的成效。同时，也有医疗机构借由“电子病历”提供以患者

为中心的个性化疗程建议，或是靠“电子病历”整合分散的数据源，大幅改善了营运绩效，让我们的医疗体系可以帮助更多人，也让所有人活得更健康、没有负担。

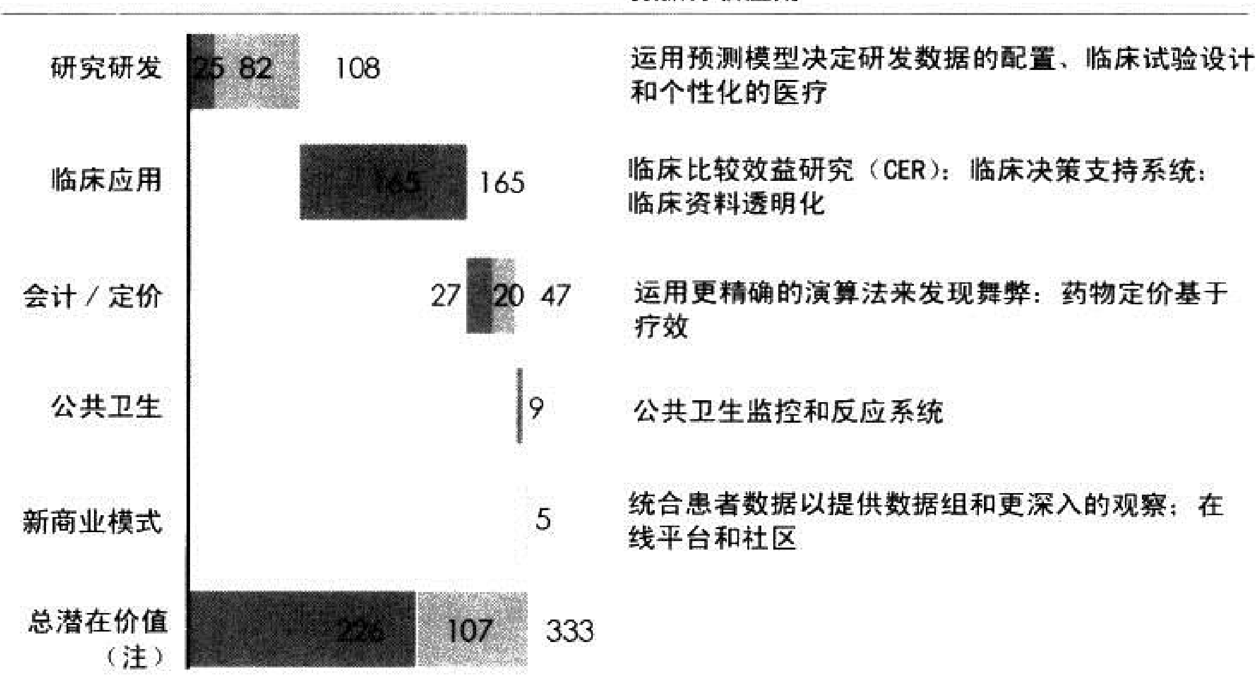
图 6-1　大数据分析对医疗业的价值

预计运用已知数据分析方法，长期可创造超过 3000 亿美元的价值，且可能节省超过 2000 亿美元的国家医疗保健支出

运用大数据的潜在价值
10 亿美元 / 年

直接减少国家医疗保健支出
对国家医疗保健支出的影响不明

数据分析应用



注：不包含最初的 IT 投资（约 1200～2000 亿美元）以及营运成本（约 200 亿美元 /年）
数据来源：专家访问、报章与文献研究、麦肯锡全球研究院

图 6-2　医疗业可应用大数据的范畴

范畴	应用
临床试验（Clinical operations）	1. 比较效果研究（Comparative effectiveness research） 2. 临床决策支持系统（Clinical decision support systems） 3. 远程病人检测（Remote patient monitoring） 4. 患者基本数据和主诉高级分析（Advanced analytics） 5. 加强医学数据透明度（Medical data transparency）
支付 / 定价（Payment/pricing）	1. 自动化系统（Automated systems） 2. 健康经济学（Health Economics） 3. 实证研究（Outcomes Research） 4. 以绩效为基础的定价规划（Performance-based pricing plans）
研发（R&D）	1. 预测模型（Predictive modeling） 2. 利用统计工具改善临床试验设计（Clinical trial design） 3. 分析临床试验数据（Clinical trial data） 4. 个人化医学（Personalized medicine） 5. 分析疾病发作模式（Disease patterns）
新商业模式（New business models）	1. 整合病历和医疗求偿记录（Claims datasets） 2. 医疗网络平台和社区（Online platforms and communities）
公共健康（Public health）	

数据来源：麦肯锡全球研究院

临床试验

所谓的早产儿是指怀孕不到 37 周就提早出世的宝宝。这些提早降临人世的小仙子，如果出生后体重不到 1500 克，很可能会因为免疫系统尚未发育完全而受到感染，一旦感染之后就很容易引发呼吸衰竭、肺出血及败血症。

不过，加拿大多伦多市立儿童医院里的早产儿，却可以睡得特别安详，因为他们是有数据资料保护的“Data Baby”（数据宝贝）。



随着医疗设备的发展，利用医疗监测仪器监测患者的生命体征（Vital signs），如血压、心跳和体温等，已经是非常普遍的事了。通常这些仪器还具有警报功能，一旦生理性的数据数值超出正常范围时就会发出警示，医护人员就会采取应对行动。但是，即使医术再精湛、经验再丰富的医护人员，可能也无法准确地察觉这些异常的发生时间和严重性，尤其当发生在脆弱的早产儿身上时。

根据美国维吉尼亚大学追踪以往的数据显示，新生儿受到感染初期的 12～24 小时，因为脉搏和心跳几乎都在可接受的范围内，因此医护人员很难从生命体征数据的改变中察觉；等到警示灯响起，常常为时已晚。

连续监测和记录这些生理性数据，可以观察出新生儿是否遭受感染的早期征兆，但是这些数据资料量实在太大了！估计这些监测设备每一秒钟会产生 1000 个读数。以往是每 30～60 分钟由医护人员归纳出一个数据做为记录，然后储存 72 小时。如果要把这些读数统统记录起来，根本是不可能的事。

但是这项不可能的任务，并没有吓跑安大略省理工学院和 IBM。他们使用来自华生研究中心的最新技术，利用流计算平台来支持大量数据的收集和分析，一天 24 小时不间断地收集和记录着包括早产儿的体温、心跳、血氧饱和浓度和血压等电子监测仪器产生的大量读数，以及周遭环境如温度、湿度等相关数据。

在保护病人隐私安全考虑下，这些数据会直接传送到安大略省理工学院研究中心和 IBM 华生研究中心；系统会分析和研究哪些因素的交互作用会造成感染，甚至哪几床的新生儿因为符合条件较多，可能出现疾

病或感染的风险较大。之后，系统再将分析结果提供给医护人员比较判读。这些动作都在数秒内完成。借由这项计划，儿童病房里的医护人员已经可以提前18～24小时，预防新生儿败血症的发生。

目前，这套结合医疗仪器和流计算的系统，还扩大应用到连结远程传感器和无线通信，可以协助系统监测病人在医院外的情况。例如白血病儿童不管是在家里、上学、或是参加体育活动，系统都可以通过和其身体连结的传感器，提早得知他们碰撞、受伤等可能导致生命危险的情况。

另外，加州太平洋医疗中心（California Pacific Medical Center）也利用大数据分析改善了心脏病的治疗效果。这个隶属萨特医疗集团（Sutter Health）的学术医疗中心，提供最专业和先进的治疗与服务，而心脏病研究是这里的重要工作之一。这个医疗中心利用数据分析系统，将本地医疗中心的临床试验数据、整个集团的患者数据库以及其他医疗机构的心脏病医学图像等不同格式的巨量数据，整合成单一平台。它还使用分析软件开发出准确的心脏病风险预测模型，建立并发症兆风险的预测诊疗模型，并让这个诊疗模型成为看诊的标准化流程之一，让心脏患者在紧急救助上所需的时间仅为58分钟，低于国际标准的90分钟；而转诊治疗的时间也由3年前的180分钟缩短到79分钟。

医学研发

位于水牛城的纽约州立大学（SUNY）是一个领先全球的多发性硬化



症（MS）研究中心。MS 是一种具破坏性的、慢性的神经系统疾病，影响全球近百万人。这种疾病会使人的大脑和脊髓发炎并产生神经病，导致患者可能会出现行动不便，视力受损，疼痛等症状。

MS 的病因是很复杂的，没有一个单一基因是可能的致病源。因此自 2007 年以来，SUNY 就一直希望从扫描 MS 患者的基因组的变化来开发新的治疗方法，通过从原本成千上万的基因序列的变异（SNP[1]）获得单一样品，研究基因产物（Gene products，例如蛋白质）和其他基因产物及环境因素进行的交互作用。

研究人员的想法是以多个 SNP 变异点结合不同的环境变因，并使用一种被称之为“AMBIENCE”的算法，来检测线性和非线性两种数据资料中的相关性，以识别这些交互作用之间的关系。但是这个想法就像是在大海里捞针，因为环境变因包括像是实验对象曝晒阳光的时间长短和维生素 D 产生的量、艾伯斯坦巴尔病毒（Epstein-Barr virus，一种常见的泡疹病毒）感染的情况、甚至是吸烟都有可能影响研究结果。况且人类的基因组由 23 对染色体组成，其中包含约 30 亿个 DNA 碱基对；这项因变量和应变量数量都大到吓人的工作，必须靠建构一套计算量可以高达 10 的 18 次方（Quintillions，18 个 0）的高等分析模型才能解决。

在没有系统可以处理这么庞大的数据量之下，以往该研究小组都是

[1] 此处指的是单核苷酸多态性（Single Nucleotide Polymorphisms），也就是人类基因组中，不同个体间基因序列产生的主要变异，常用于研究家族间的遗传现象及估测罹患疾病的难易度。

以寻找“最少数量”的SNP、环境和形态（Phenotypic）的变量组合进行研究，直到他们和IBM合作，建立了一套搭配软硬件的数据分析系统，以往平均需要27.2小时的工作，缩短到现在只要11.7分钟即可完成。而且这套系统不仅大大简化和加快了复杂的分析过程，而且还提供了不同类型的变量值，例如分类变量、分立卜瓦松变量或连续常态分布变量等。过去，只要在研究中增加一个新的变量值，研究团队就必须重新编写整个算法，而现在却只要按几个按键就可以达成。

大数据分析系统除了应用在MS研究以外，全球估计有超过3300万人感染、至今还没有方法可以完全治愈的艾滋病，以及罕见的遗传性骨疾病，都已经开始利用这个新技术进行大型的医学研究。科学家们不用再花大把时间去思考如何编写算法，而是可以把精力专注在医学研究上，来促进全世界人类的健康。

电子病历

1933年创建的广东省中医院是中国历史最悠久的中医院之一，更是中国南部最大的医院系统，目前旗下拥有5间医院，院内有超过3000张病床，每年门诊量超过600万人次。这家一日门诊量就能破万的医疗机构，之所以备受推崇，是在于它独特的治疗技术，广东中医院主要是以结合中药和西药，双管齐下的方式给予患者治疗。

不过，也由于服务的患者众多，2008年以前，包括病历或是临床试验等数据都由各家分院的各个单位自行保存；也因为独立记录管理，这



些医疗信息无法在部门或分院间共享。为了提高服务水平、打造一个以患者信息为中心的电子病历系统，广东省中医院需要将以前分散在多个系统中的医疗数据，整合成为一个标准化规格并可以进行分享的数据中心。

广东省中医院和 IBM 携手合作了一套“临床记录分析和共享”（CHAS）系统，可以把这些横跨中西医的数据都整合成以单一患者为中心的标准电子病历（EMR）。这套系统不仅把传统中医上的理念和智慧，转变成可以严格、高效进行测量的标准化数据之外，也使用语义分析技术，结合了中西医两种智慧，使得不管是哪一种格式或语言的中西医学术名词都有详细解释，让医护人员可依此向患者说明。

之后，广东省中医院和 IBM 继续合作建立“医疗数据库分析和共享”（HIWAS）系统，储存并整合匿名病人的数据，其中包含年龄、性别、是否患有如心脏病或糖尿病等其他病症的详尽数据。系统并在诊疗过程中帮助医生取得、过滤并整合其他相关病人的数据和类似的医疗行为，协助医生为病人量身订制个性化的治疗方案。

此系统也被运用在慢性肾脏病的治疗。目前，引发慢性肾脏病的最主要原因之一是高血压和糖尿病等文明病。根据统计，约有 4 成的糖尿患者会成为慢性肾脏患者；美国成年人中就有 17%罹患慢性肾脏病，而广东中医院每年也需要治疗高达一万名的慢性肾脏患者。

然而，这一类的患者虽然人数众多，但是着眼于该疾病的西药疗法十分少见。广东中医院针对上千份的匿名电子病历进行搜集并分析，找出中西医对于慢性疾病治疗的关键信息，并从中归纳出不同群体对慢性

肾脏病的发病规律，目前已具有相当的疗效。现在，广东省中医院不仅可以比以往多治疗25%的病人，并且获得多出10倍的临床信息，往实证医学和更具成本效益的医疗发展又向前跨了一大步。

改善营运

Premier Healthcare Alliance 是一个国家级的医疗采购集团，与美国的全国联邦医疗保险（Centers for Medicare & Medicaid Services）和英国国民健康制度（National Health Service）都有合作。目前该医疗联盟在全美拥有超过2600家会员医院和86 000家其他医疗机构，希望借由协作的力量提高会员医院的运作效率及成本效益，同时帮助他们适应责任医疗（accountable care）和预计2014年全面实施的新型保健支付及求偿（delivery and reimbursement）等医疗改革。

由于会员涵盖面极广，Premier的临床数据、实证研究和财务营运等数据也非常丰富。但是他们却发现，现有的IT基础设施无法应对即将实施的新型医疗保健制度，于是他们花了一年多建立许多数据储存系统。但最后却发现这么做的结果是，会员根本无法串连不同来源的数据，并且因这个系统缺乏扩充性，也无法支持持续大量增生的数据量。

要怎么顾及这些数据的隐私和安全，甚至在机密文件上可以防止有人越权存取，又能够让成千上万的组织会员，可以共享医学数据和研究成果？Premier的构想是为每一个医疗机构的数据，设置独立的安全保护，但同时又使之能做某个限度的分享。利用IBM的数据分析系统和社交网



络平台，会员们不仅可以利用社交入口网站得到各种临床、业务和规范等相关数据，还可以借由其他大型医院或研究单位的临床结果，改善病人的治疗状况。

医院经营者可以利用这个系统比较国家制定的标准和自己的营运效益，以帮助他们在医疗保健服务和行政成本上提高资源利用率、减少浪费。而医护人员则是能够从大量病历集结而成的数据库中，找到最适的治疗方法。例如，若一名糖尿患者受了外伤而进入医院，系统会将此患者的所有数据整合在一起，使这些数据规格化、基准化，并且比对数据库中所有类似情况的病历，为医生或药剂师提供抗凝血药物的适当剂量。

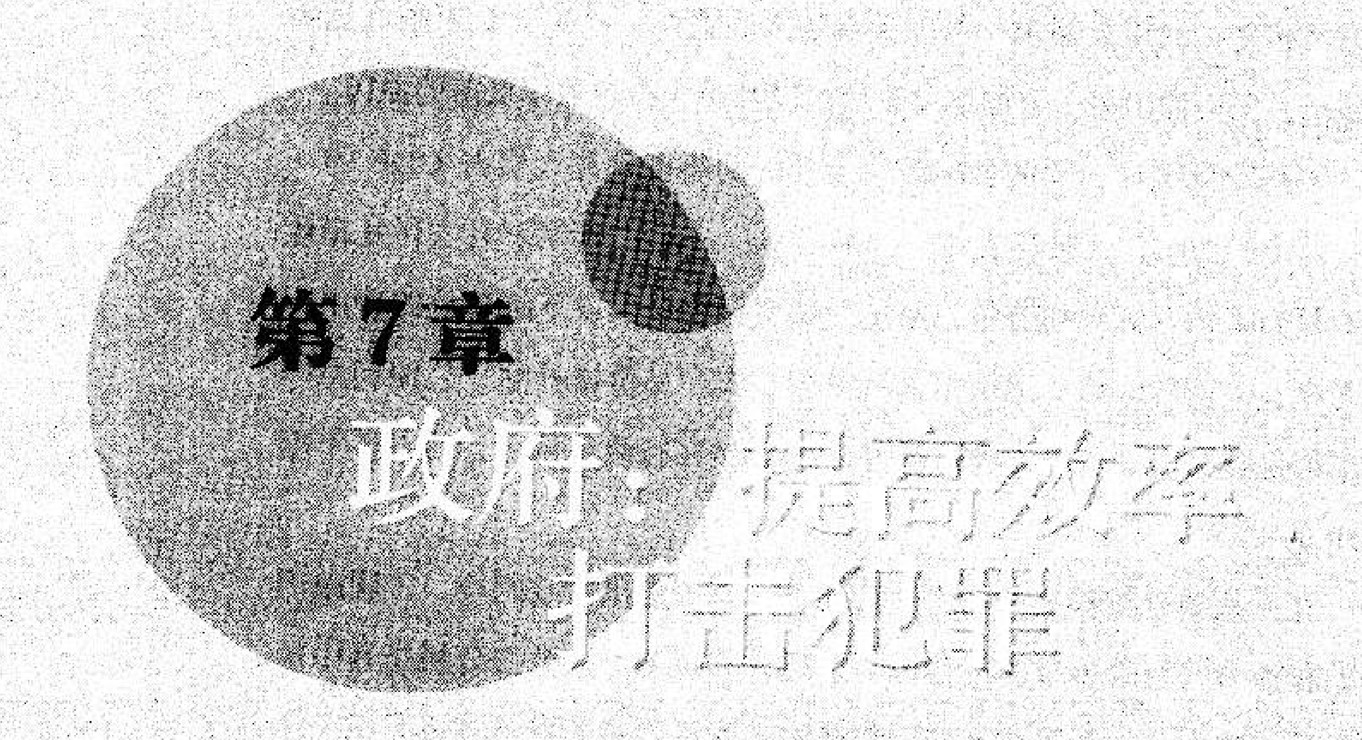
同时，为了让非会员的医疗机构也可以改善营运、提高效益，Premier 也和 IBM 合作建立一个包括临床试验、操作过程和实证结果的数据模型做为范本。这个被称为“医疗服务提供商数据仓库”（IBM Healthcare Provider Data Warehouse）的系统，可以帮助所有医疗机构收集和分析准确和实时的信息，提供以证据为基础、以患者为中心的责任医疗服务。现在，从入院、出院和转院（ADT）的患者数据，到如制药、微生物学和实验室等来自各医院、各部门的信息，都会被发送到核心数据库中，以每秒 3000 TPS（Transactions per second，意指每秒可处理的事务数）的速度处理，不仅让研究成果和临床诊疗可以更快、更容易地结合，同时也降低了营运成本。

这个方案实施后，估计已为 Premier 旗下会员减少了 28.5 亿美元的医疗开支，并挽救了 24 800 人的生命。目前，Premier 已计划，利用新科技和新思维，设计更多的新应用程序以支持医疗改革和其他立法要求，

因为它知道，这样的愿景不是在明天，或是三五年后的未来，而是现在就要开始！

现在中国台湾地区正面临着医疗保健巨额亏损。根据台湾省的最新统计，和10年前相比，台湾人在教育、休闲文化、衣着鞋袜的支出减少了，医疗保健的支出则是明显增加了3.2%。这和中国人均寿命延长以及人口老龄化有关。同时，台湾医疗机构的护士荒以及内、外、妇、儿与急诊科医师严重出走的问题也越演越烈，看了上述案例，你是否心有所感？医疗机构如何运用新科技、新技术增加生产力和降低成本，或许这才是当前医疗体系所需要的处方。





第7章

政府：提高效率、打击犯罪



改善公共行政效率是许多国家政府正在面临的巨大压力，尤其是在这几年全球经济衰退的情况下，政府一方面必须维持甚至增加公共服务，另一方面却得面对因为财政赤字所带来的预算限制。

尤其是在欧洲，估计到 2025 年，欧洲国家 60 岁以上的人口，将占总人口数的 30%。也就是说，届时每 3 个人当中就有一位是 60 岁以上的老人。这些老龄人口的社会福利、健康问题和退休金，所延伸出来的是对医药和看护的庞大需求，需要大量的人力资源；结果人数不断增加的公务员，又衍伸出另一个加重财政负担以及造成行政效率低落的原因。

根据欧盟统计，欧洲各国政府公部门的年度预算大概占 GDP 的一半，其中的 20%用于转移支付（transfer payment）。所谓移转支付是指政府不是为了购买商品和劳务，而是无偿支付给个人或下级单位的费用；养老金、失业救济金、退伍军人补助金、农产品价格补贴、公债利息等支出都包含在内。剩下的预算则有 30%花费在行政单位上，约占整年度 GDP 的 10%。

在无法减少转移支付费用的情况下，为了走出财政困境，各国政府开始积极节约行政单位的运作成本，提升其效率。大数据分析是其中最有效的方式之一，几乎所有国家都能因此受惠。然而目前却不是每一个国家都已经开始进行，其原因在于人为政策的干预。

虽然，政府所产生的数据大多是文本和数字，而且有 90%以上是数字数据，但是这并不代表它比企业或个人的数据更有效率地被运用。最主要是因为各单位并没有一致化的标准，且碍于各自的政策、规范，无法完全共享数据。所以，A 单位必须用传真的方式递送公文，或是邮寄

一张光盘数据给 B 单位的情况，到现在都还很常见。数据缺乏透明度的结果，也让政府部门难以制定促进绩效和生产力的指针。

所幸，有越来越多的政府开始采用“开放数据”（Open data）的原则，把不受著作权、专利权，以及其他管理机制所限制，经过挑选与许可的原始数据，开放给社会公众，让任何人都可以自由运用。英国的 Data.gov.uk 和西班牙的 www.proyectoaporta.es 就是其中的先锋。

这两个国家的政府已经有了共识，认为此做法除了让全民可以监督政府部门的效率之外，也可以利用数据分析、测量和比较不同单位的职能，建立内部的竞争驱动性以提高生产力，甚至公务人员也可以利用这种机制衡量供货商的表现，以减少错误和舞弊。MGI 估算因此将可节省高达 30%的政府采购成本，甚至减少 3%的转移支付，因为转移支付中错误和舞弊的金额非常高。

以目前最需节约成本的欧盟 27 个会员国来看，麦肯锡全球研究院估计借由大数据分析，OECD 中的欧洲国家政府可减少 15%～20%的行政成本，并在未来 10 年间每年增加 0.5%的生产力，换算成量化数值每年约为 1500 亿～3000 亿欧元（见图 7-1）。此外，大数据还可以提供一系列无法量化的价值，包括预算分配的改善、更高质量的服务、加强对公共部门的问责制（Accountability System），而这些都会增强民众对政府的信任。

除了提升公共部门的行政效率之外，各国政府也开始将数据分析应用在协助打击和预防犯罪、解决城市交通拥塞以及防治污染上，希望借由更新的技术，取代更多的人力，而且在确保民众的隐私安全之下，发展更好的决策、建立更健全的政府。



图 7-1 欧盟国借由大数据分析锁获得的潜在效益

OECD 欧洲国家（OECD-Europe）公共部门可借由大数据分析创造 1500 亿到 3000 亿欧元（甚至更多）的潜在价值

		总计数（注1）10 亿欧元	X 可锁定 %	X 可减少 %	= 总价值 10 亿欧元
提高行政效率	行政成本	4 000	20～25	15～20	120～200
减少错误和舞弊	转移支付	2 500	1～3（注2）	30～40	7～30
增加税收	税收	5 400	5～10（注3）	10～20	25～110
					150～300+

注：

1. 提高行政效率带来的价值，基数不包含转移支付的总政府支出；减少错误舞弊所带来的价值，基数是总政府转移支付；增加税收带来的价值，基数是税收。
2. 包含转移支付中可能有的错误舞弊及其预计将造成的损失。
3. 税收部分，可锁定的百分比是指预计税收缺口所占百分比。

数据来源：国际货币基金；OECD；麦肯锡全球研究院

提高行政效率

Age UK 是英国最大和最知名的慈善机构之一。这个成立于 2009 年的慈善机构，专门关注老年人的权利、生活、安全和保健等议题，每年提拨 7500 万英镑的资金防止虐待老人，并为贫穷或有需要的老年人提供医疗保健和社会服务，每一年的受援人数超过 500 万人。

“间接成本”是该机构最大的敌人。为了争取慈善捐款，他们必须将预算尽可能有效地放在营销资源的使用上，而包括邮寄、办公室等其他间接成本都要尽量节省精算，因为他们花在这些支持募捐活动运作上的成本越少，可以运用的捐款就越多。

不过，Age UK 不管怎么节省，庞大的邮寄费用都始终无法下降。以往，Age UK 会针对每一次的活动，将广告邮件寄给过去 4 年里曾经捐款的所有捐助者。他们虽然想减少邮寄广告 DM 的费用，但是却不想冒着因为省钱而失去可能得到高额捐助的风险，所以不管是捐了 5 英镑或 50 英镑的捐助者，全都收到了邮件。

Age UK 希望着眼于捐了 50 英镑的高额捐助者，以确保大部分的筹款收入，但又同时想减少邮寄费用。因此他们积极寻找一套可以纳入更多变量的解决方案，让他们可以更深入地了解每一个支持者所提供的捐款，再配合个人的邮寄费用进行分析，以列出最具效益的邮寄名单。

这个想法说起来简单，但实行起来并不容易，因为对 Age UK 来说，每一笔捐款都很宝贵，而且十几个小额捐助者和一个大额捐助者，对于邮寄成本的影响到底怎么计算？如何才能以更少的成本，吸引更多的大额捐助者？最后他们采用了 IBM 的数据分析系统，针对每一个捐赠者的单次捐赠金额、捐赠的频率和时序加以统计分析。

筛选的逻辑是依照商业上的“近因频率的货币价值”（RFM），衡量顾客的 3 种特征值而定，包括了最近购买日（Recency）、购买频率（Frequency）及购买金额（Monetary）。首先系统将现有的百万名捐助者，依照以上 3 项特征进行排名，再区分为 5 个族群，然后再依设定条件为



每一个捐助者评分，以判断寄信给这位捐助者是否符合这一次活动的邮寄成本。

现在，Age UK 已经利用这套数据分析系统，成功削减了邮寄成本，但仍然保留了维持捐款额度的主要支持者，而且每一个列在直接寄送邮件名单上的捐助人，响应速度和总贡献度都增加了 100%。不仅如此，Age UK 还利用这套系统进一步细分捐助者的年龄、居住地、职业和其他人口统计变量，帮助该机构推出更多针对不同属性捐助者的宣传活动，实行后有些活动所募到的金额甚至较往年还高。

利用数据分析系统，把补助款和资源用在真正需要的人身上的，除了 Age UK 之外还有加州的阿拉米达社会服务局（Alameda County Social Services Agency，ACSSA）。

有"黄金之州"称号的加州，截至 2012 年 7 月为止，已经有 3 座城市宣布破产，还有 8 座城市正面临重大的财政难关。据统计目前在加州生活于贫穷线之下的人口已超过 600 万人，全美因缴不出房贷而房屋被依法拍卖的数量，也以加州占最大比例。加州的第 7 大郡阿拉米达更是其中之最，每个月有超过上千户家庭会收到房贷拖欠通知。

目前，阿拉米达 160 万人口当中，有超过 15 万人生活在贫穷线以下。他们的生计很大一部分是依靠 ACSSA 的补助，而该机构一年要管理的案件高达 19 000 宗。但是庞大的业务量和陈旧的系统早已无法支持其需求，导致该机构常常在事件发生数周，甚至数月后才获悉，不仅处理效率不彰，而且可能导致资源浪费和舞弊行为。

ACSSA 意识到他们需要在社工提出要求时，及时地提供更多准确的

案件相关信息。2008 年 ACSSA 得知 IBM 为服务旧金山的寄养儿童开发了一种系统，可以追踪该系统记录在案的人，而且还具有分析的功能，可以帮助工作人员厘清个案情况的因果关系，运用成效相当良好。

因此，ACSSA 决定导入这种结合分析功能和商业智能的系统，经过 2.1 个月的导入期后，这套系统不仅可以实时为个案社工提供项目的详尽信息，让他们为每一种情况找到最佳的辅助方案，而且还能实时追踪受益人与个案之间的关系，有效防止舞弊和冗余的现象。另外，以前需要数周或数月才能获取的报表，现在也只需几分钟时间就可以得到。这也使得社工员有更多时间在现有数据基础上，利用“假设方案”（what if）模拟对于个案更好的处理方法。

目前这套数据分析系统已经提高了为该机构辩护的律师的效率和获胜率，每年约节省了 90 万美元，并减少资源浪费超过 1 100 万美元，整体效益的年平均值接近 2500 万美元，换算成投资回报率（ROI）高达 631%。

打击和预防犯罪

先前提过最早将数据分析运用在公共安全管理上的纽约市警局，借由分析犯罪数据的时空分布形态，进而预测犯罪机率最高的时间与地点，配合警察巡逻的路线调整，让重大犯罪事件降低 30%，更让纽约从罪恶之城变成全美最安全的城市之一（America's Safest Cities），CompStat 系统也因此成为全世界警局争相仿效的对象。

有了纽约市警局的成功案例之后，包括孟菲斯、纽约、查尔斯顿和



南卡罗莱纳州的警察局都和 IBM 合作建立警用数据分析系统，利用过去犯罪活动的数据库，以及时间和天气等相关消息进行犯罪活动预测，并呈现在地图上。

虽然数据分析系统的成效卓著，但在实行上并不如想象中容易，因为绝大多数的警察和连续剧《CSI 犯罪现场》里善于利用高科技办案的形象相距甚远。在使用新技术方面，美国许多警察局甚至落后于青少年：一些警车仍在使用磁带式录像机，或者是使用复写纸写报告。另一方面，各个州政府目前都在为紧缩的财政焦头烂额，警察的预算也变得相当吃紧。

孟菲斯是一个典型的南方大城，人民热情而保守，2004 年就任的孟菲斯警察局局长鲍德温（Larry Godwin）也一样。在担任局长之前，他是一位在海军陆战队服役 38 年的军官，受过非常严格且传统的训练，退役后转任警界，负责管理辖下 9 个分局、2470 名警员。

然而一上任，鲍德温就面临犯罪率不断攀升，而预算却在缩减的窘境。他必须要在最短的时间内，拿出最有效的方式，保障 67 万市民的安全。为了有效降低犯罪率，他在这个作风保守的地区，大胆采用了 IBM 数据分析系统，制定了蓝色粉碎（Blue CRUSH，Criminal Reduction Utilizing Statistical History）计划，初期却遭致议论，因为警察们都认为应该另外增加 500 名巡逻警力，以抑制增长的犯罪活动，而非花钱买科技设备。

但是鲍德温知道，调整现有的警力资源，使警察办案更有效率才是根本解决之道，因为警察们需要的是更有智慧、而不是更辛苦地办案。

鲍德温在获得市长支持之后几小时内就开始执行这个计划。他向警察传达两个重要的消息：第一、所有的报告和数据都要采用标准化规格；第二、每周举行的各区指挥官会议改为 TRAC（Tracking for Responsibility, Accountability and Credibility）会议，追踪各单位的实行成果。这两项命令没有模糊空间，而且立刻执行。

同时，数据分析系统也开始比对历史数据和最新的街头监视器画面，并深入了解促使犯罪的长期因素（例如废弃房屋的周边），了解当地不同犯罪类型的趋势、发生的地点时间等，计划实施 5 年以来，孟菲斯的犯罪率已降低了 30%以上，暴力犯罪降低了 15%，重案组的结案率更是从 16%大幅提高到 70%。

瑞士日内瓦警局也遭遇类似的情况。这个在国际上享有高知名度的城市，有许多全世界最重要的国际组织在此设立总部，包括红十字会、世界卫生组织和联合国欧洲总部等。

曾经荣登全球最佳居住城市的它，近年来犯罪率却上升了超过 15%，2010 年发生在日内瓦境内的罪行就高达 61 910 件，高于全瑞士的平均水平。对当地警察局来说，这是一个非常令人担忧的趋势，但是更令人忧虑的是因为政府预算，日内瓦警局的人力必须缩减 3%，在这种情况下该怎么积极和有效地打击犯罪？

为了降低犯罪率，日内瓦州警察局利用数据分析技术加上制图工具，将这些空间数据和非空间数据的分析结果整合在一个平台中，每天早上警方可以收到一份不断更新信息的犯罪报告，分析当前的犯罪活动、犯罪模式，并有助于预测未来可能会发生的罪行。



另外，通过互动地图和系统上装置的仪表板，每个指挥官都可以随时查看各地区发生的案件，如抢劫或意外事故等，不用等到警局通知，就可以知道哪里应该加强部署人力，或是在哪些地区加强巡逻。

例如，将抢劫、盗窃、袭击或是毒品交易的案件数据，依照地理位置进行编码，经过分析后，系统发现青少年犯罪的周期，大多集中在星期三下午、星期五和六晚上，且发生地点在距离市中心最远的街区内。警方可以依此决定该采取什么样的行动，以阻止或预防犯罪活动，并在正确的时间和地点加派警力巡逻。自从这个解决方案推行以来，日内瓦因此减少了 3%、相当于超过 1800 件的罪行，也重拾了世界最佳居住城市的荣耀。

除了警务工作之外，这几年来激进份子的恐怖袭击，更使数据分析技术在公共安全管理上有了重大突破。美国 9.11 事件爆发后的 911 天，恐怖份子在西班牙马德里引起了另一次震惊世界的连环爆炸案。

2004 年 3 月 11 日早上 7 点半，正是人们的上班高峰时间，西班牙马德里市郊的短途火车一如往常地载送乘客，没想到 10 分钟后，列车正要驶进马德里阿托查火车站时，轰地一声，车厢被炸得面目全非。与此同时，在开往阿托查火车站的铁路线上，包括市中心附近的蒂奥雷蒙多火车站和圣欧亨尼娅火车站也相继发生爆炸。

这次袭击，在 4 列火车上共发生了 10 次爆炸；炸弹由手机引爆，而引爆时间设为当地上午 7:39 分。13 个土制炸弹中，有 10 个成功被引爆，造成一共 200 名死难者、2000 多名伤者。在这起举国震惊的爆炸案发生后，马德里市议会成立了紧急应变中央指挥中心（CISEM），加强城市保

护机制。以往，警察局、消防局、救护车中心、部队、和巡警在处置突发事件时，因为各自有自己的指挥系统，每个部门还有各自的通信系统和相关技术设备，所以难以进行任何形式的联合行动方案。

作为一个中央指挥中心，CISEM 目的就是要让各单位在突发事件发生时可以马上联合运作，并且减少各单位的标准流程和响应时间，同时提供准确无误的指令。IBM 于 2006 年开始与 CISEM 一起整合马德里各个安全管理单位的通信和响应能力。首先，每辆警察巡逻车、救护车和消防车都配备了掌上电脑（PDA），用于通信和接收来自 CISEM 中央指挥中心的指示。

然后，IBM 和 CISEM 整合了不同部门使用的所有应用程序，和包括马德里 112、市紧急热线、影像监控中心、M30 高速公路控制中心等外部单位的 IT 基础设施，形成了一套中央应用程序。而这一套程序也用以管理在马德里举办的活动，像是足球赛等大型公众集会。各单位通过移动设备取得数据和接收命令后，可以更有效率地布署警力和救护车。

第 3 步则是数据整合，将各单位的信息全都整合至一个中央数据库。只要市民遇到紧急情况并拨打 112，系统就可以自动同时通知警察局、救护车服务或消防局。就算从多个不同来源同时收到紧急事件的通报，因为这些信息被清楚地标示是属于同一起事件或是一系列不同事件，便可避免人力的重复。而且，各单位可根据不同的紧急事件分配最适当的资源，从而使资源调动更快、更有效。自从有效整合之后，马德里发生紧急应变事故的次数已减少了 25%。



改善交通问题

瑞典斯德哥尔摩，是一座由岛屿组成的城市。14 个市镇分布在大小的岛屿上，由各种桥梁相连，居民们大多习惯驾车穿行于岛屿之间。

这一天，和往常的冬日下午一样，黑夜早早降临，下午 4 点钟就到了斯德哥尔摩的交通尖峰期。汽车在路上呼啸而过，顺畅的车流代表着一个先进的城市。数年前，斯德哥尔摩和世界上任何一个大城一样，车多路少，道路建设满足不了交通需求；市政当局鼓励人们乘坐公共交通工具，但拥堵情况仍越演越烈。

每天都有超过 50 万辆汽车在这座城市中穿梭，2005 年时人们上下班的通勤时间已经比前一年增加了 18%。为了解决交通堵塞，瑞典国家公路管理局（SNRA）和斯德哥尔摩市政府在 2006 年初宣布试征“道路堵塞税”。这套道路收费系统与新加坡、伦敦和奥斯陆等城市实施的政策相似，根据每一天不同的时间对经过的车辆收取 10、15 或 20 的瑞典克朗；最高收费是在上午 7:30～8:29 和下午 4:00～5:29 的尖峰时段。

面临群众的反对声浪，市政府提案：先进行为期 6 个月的测试，之后再举行市民投票，决定该项目施行与否。碍于压力，政府官员要求这个系统必须在 96 公里的时速下，“扫描”将近 50 万辆汽车，并且精确识别超过 99%的所有车辆。

原本系统的规划是在每个车道的上方都放置两个摄影镜头，朝向相反的方向，分别捕捉前后车牌，并且将摄影镜头的视野延伸至邻近车道，

以便从各个有利的角度拍摄每辆汽车。但一开始识别的成功率只有60%。后来IBM开发出光学字符识别软件（OCR），它可以从任意角度辨别车辆的牌照，并且针对光线强度的不同、天气恶劣或拍摄视角欠佳、标准系统可能无法识别车牌号码，利用算法进行二次识别。

这些算法使用的技术，是模仿人类的眼睛，例如增强图像和比较车子头尾的车牌分析，直至获得最佳的视角。通过这项技术，道路收费系统才顺利达到了99%的精确度，而且每天经过的近50万辆汽车中只有4到5辆会漏失拍摄，执行方案也才获得市政府的认可。

为了在不增加人力的情况下顺利收费，IBM 在斯德哥尔摩城区出入口，规划了18个车主无需停车的路边收费系统。车辆通过第1道激光束时，会触动收发器天线，使收发器向车辆的车载传感器发出信号，并记录下当下的时间、日期和应缴税额。在收发器工作的同时，摄像头会拍摄车辆的车头牌照，等到车辆通过第2道激光束时，第2台摄像头拍摄车尾牌照。以上所有步骤都可以在毫秒之内完成，车辆无需减速。之后，费用将从车主的账户扣除，或是驾驶者也可以通过网络、银行以及7-Eleven等便利商店支付。

实施后，这套道路收费系统果然缓解了斯德哥尔摩的交通堵塞情况，市中心的交通流量锐减25%。而且，因为通行时间缩短，公交车公司不得不重新设计交通时刻表，而使用斯德哥尔摩公共交通工具的平均人数，也比前一年增加4万多人。甚至，因为车流量减少，道路的废气排放量减少了8%～14%，市中心的二氧化碳等温室气体排放量也下降了40%。

此外，试行期间，为了怕民众投诉，市政府还设置了电话客服中心，



税务局甚至雇用了40名律师，处理可能发生的上诉案件，但投诉电话却没响过。最后全体市民经过投票，通过了这项交通方案；他们都同意，数据分析系统让斯德哥尔摩回复了原本的生活质量，甚至比原来的更好。

斯德哥尔摩交通整治的成功案例，让其他有交通问题的城市找到了新希望，例如澳洲的昆士兰。

昆士兰是澳大利亚人口增长最快的州，每星期有超过1500人移居该州，其首府布里斯班市的劳工人口数量预计在未来20年几近翻倍。爆炸性的人口增长，造成交通拥塞的情况越来越严重。2005年州政府和昆士兰高速公路有限公司（Queensland Motorways，QML）合作投资18.8亿澳币，进行盖特韦大桥（Gateway Bridge）升级改造和配套的道路工程；2007年扩大项目规模，延伸总长61千米的洛根（Logan）和盖特韦延长高速公路（Gateway Extension），而且道路容量倍增至12个车道。

这个道路系统规划，以多车道高速公路由南蜿蜒至布里斯班市北部，并绕过中央商务区，为前往布里斯班的港口和机场的商务人士提供便利运输方式。而且该路线中途也穿越了布里斯班河上最具代表的盖特韦大桥，成为昆士兰交通基础设施最重要的部分。

不过，兴建过程中州政府和QML同时意识到，不断地增建道路并不能永远改善交通拥塞的问题，因此，除了实体的基础设施投资之外，还需要其他可持续发展的交通缓解政策。在研究如何转变交通管理流程时，州政府和QML发现车辆在通过收费站时，不管是放慢速度使用电子系统付费，还是停下来用现金付款，车流速度都会大大降低，使得收费站附近的车流更加拥挤。

因此，如果能够使收费流程自动化且无需驾驶停车，将可以立即提高车流平均速度，同时若能数字化采集并分析关于道路上车辆的数据，也有助于日后改善交通管理。最后，他们决定和 IBM 合作于 2011 年完成收费系统的完全电子化。

这套名为自由畅通收费（Free-Flow Tolling）的系统，以高科技的路边吊架取代传统的收费亭，利用摄录像机和专用的短程通信技术追取过往车辆的数据，再通过车内条形码或使用 OCR 系统分析车牌号码镜头来识别车辆。随后，车辆数据传送到相符的民众账户，由自动化计费系统评估车主需要缴纳的费用，并从客户账户的预付费用中扣除。

实施之后盖特韦和洛根高速公路的车道通行能力，从每小时约 300 辆车辆，提高到了每小时超过 2000 辆车辆，交通时间也缩减了 10 分钟，甚至因为车流顺畅，高速公路上的交通事故也减少了 27%。

此外，运用这些数据，QML 可以了解路上行驶的车辆类型以及这些车辆上路的频率和时段，并且结合这些通勤交通模式数据和实时路况数据，未来可为车主打造量身订做的路网规划和建议。例如，每周一早上前往机场的车主可以直接在手机上收到交通拥堵情况的相关报告，并参考其最佳路线规划。

这个预测交通的智能功能目前已成功在新加坡应用。人们能像获得天气预报一样，获得交通堵塞预报。通过埋在路上的传感器和红绿灯上的探头，司机不仅可以看到什么地方在堵车，还能够提前预测什么地方在 10～20 分钟内会堵车，从而选择更为通畅的道路行驶。美国波士顿也正在跟进，制定类似的交通整治计划。因为运用了数据分析技术，城市



的交通，将走向更加畅通无阻的未来。

防治污染

请不要让你的眼泪，成为地球上的最后一滴水。

这是几年前，一句发人深省的公益广告词。触动人心的背后其实隐藏了一个更大的危机。根据联合国研究指出，气候变迁和人口快速增长导致的粮食、能源和卫生需求增加，已经使得全球水资源供应趋紧，加上过去50年抽取地下水的数量增长2倍，许多地区的水供应已经因为降雨型态改变、严重干旱、冰川融化和河川流量改变而减少，今天全球已有1/5的人不能得到足够的安全饮用水，到2080年恐将有半数人口面临严重的水资源短缺问题。

联合国还预估2020年至2050年间，处理水资源问题每年需耗资137亿～192亿美元，而水资源的枯竭和污染也成为各国政府关注的首要议题。

位于美国纽约的哈德逊河全长约520公里，河的两岸孕育着近1200万人口，文化遗产和自然生态资源都非常丰富。但是二次世界大战后，许多跨国企业例如通用电气、通用汽车等都在哈德逊河沿岸设立了自己的工厂，工业的发展使得这条河在一定的程度上遭到了污染。

如何恢复哈德逊河的原有风貌成为纽约市政府的课题。市政府与IBM合作，建立了全球第一个电子化的水资源管理中心（AWM，Advanced Water Management），并在哈德逊河中放置了很多的传感器和电子卷标。

这些智能浮标可以让传感器通过无线网络送回计算中心，进行计算模式的分析和处理。所有数据会在计算机系统里被显示成一个虚拟河流，每一个层面的变化都可以实时被控制，用以协助环保政策制定、区域经济规划和水生物管理。

这套水污染防治系统，在 2008 年被移植到了爱尔兰的高威海湾（Galway Bay）。IBM 和爱尔兰的工业发展局共同合作，借助传感器汇整入海河流的流动状态、海浪高度和浮游生物、水质、鱼群等大量数据，监控高威海湾的污染水平和其他环境状况。

在像高威海湾这样的封闭水域里，污染物扩散的危险比开阔的海面产生速度更快，因此需要实时获得和解读海湾的相关信息，并且立即对任何危险信号采取防治行动。因此，研究人员在海湾上装设配有射频技术的智能浮标，并配备了传感器，用来收集包括盐度、温度、波浪能、潮汐和海湾内鱼类和植物等各种生物的状态，通过流计算 24 小时不间断地分析这些数据。

由于系统可与其他在线数据库（例如地理空间信息）结合在一起，不仅可以作为水污染预警系统，也可以让气候研究人员利用陆地上的传感器和海湾中的传感器，了解陆海接口的二氧化碳交换，测试全球气候变化的长期影响；海洋生物学家可以使用海湾中部署的声波传感器评估海洋哺乳动物数量；实时的潮汐数据也让波浪发电成为该地区的另一个替代性能源。

除了防治水污染之外，近期数据分析系统也开始用于空气质量的监测。由于全球变暖现象，使得森林火灾在近几年发生的频率越来越高，



根据联合国统计，全世界每一年受到森林和荒地火灾影响的土地高达3.5亿公顷。

无情的野火不仅给人类带来了巨大的生命和财产损失，还使得气候变化、空气污染和丧失生物多样性等问题进一步恶化。例如，2009年发生的澳洲维多利亚森林大火就导致2000多所房屋被焚毁，175人死亡，造成的经济损失高达15亿美元。2012年发生的美国科罗拉多州野火，也迫使附近居民约3.2万人仓皇逃离家园，大火灰烬直冲云霄，高达6100米，也为当地的环境生态带来一场浩劫。

马里兰大学巴尔的摩分校（UMBC）从2008年开始研究野火烟雾（wildfire smoke）的扩散模式，以提供消防及公共安全官员实时评估火势。以往，烟雾模式的分析仅限于气象预报，且大多是以低分辨率卫星图辅以一线工作人员的意见，大约每6小时才能更新一次；而使用IBM的流计算系统，分析森林大火的烟雾方向时，研究人员可以立即处理从无人驾驶飞机、高分辨率卫星图和空气质量传感器收集而来的大量数据，建立驱散烟雾的有效模式，并且随时更新。

要建立烟雾扩散的数学模式非常不容易，因为它会随着不同的风向、气候，甚至是人为因素而随时改变。同时，因每一个地方的扩散烟雾颗粒浓度不同，数值也会不断变动。因此，研究人员首先要从多个渠道，如美国航空局、国家海洋和大气管理局（NOAA）建立的空中卫星传感器，以及地面上的烟雾探测器等收集实时的空气质量数据。

不过，由于美国国家海洋和大气管理局设定的静止卫星，提供的测量大约每半小时一次，而NASA（美国国家航空和航天局）提供的轨道

卫星，测量操作一天两次，所以还需要搭配无人飞机回传的影像，以此更准确地判别空气中烟雾颗粒物质的浓度。

之后系统通过数据分析系统里的气滞扩散模型（HYSPLIT），预测烟雾颗粒在大气中扩散的走向，并以此计算出浓烟的覆盖范围，预测火和烟的走势，并建立相对应的空气质量监测虚拟模型，协助消防人员更准确地了解森林火灾的蔓延趋势，实时发布预警，以减少死亡人数和财产损失。

目前研究还在进行当中，但这是人类第一次有能力利用数据分析，对气体进行预测，且准确率较先前提高了约 16%。这不仅仅是科技带动污染防治和公共安全的第一步，更是利用人类科学方法洁净地球、回复生机的开始。





第8章

能源：节能减排新利器



高油价和高电价让可持续能源议题持续发烧，也使得大数据分析在能源产业的重要性与日俱增。最明显的例子是在第 1 章提过的美国政府“大数据研究和发展计划”。这项计划总共耗资两亿美元，其中有 1/8（2500 万美元）拨给能源部（Department of Energy），用于研究如何运用大数据加强能源使用效率。

在业界，大数据分析因为直接影响到营利，更是了不起的大事。能源业涵盖从探勘、生产和运输石油及天然气等能源的公司，到负责发电和供电的电力公司。不少企业已经装设智能化的监测设备，实时收集大量的作业数据进行仿真分析，以提高生产力、降低成本，并实时评估设备稳定度，防止运作中断或发生事故。

以油气探勘为例，多数人以为石油和天然气储藏在地下岩洞中，只要找对地方就可开采。事实上，油气大部分都是蕴藏在地下岩层的孔隙内，能源公司必须收集油气层上方每个位置的地质数据，有系统地整理，再利用已知点的数据推论未知点的特性，以尽可能描述整个油气田的动态。

由此可知，油气探勘本来就是信息非常密集的工作，再加上现在由于容易开采的传统油井日益稀少，企业只好越钻越深、并往危险性高的地区开采。在这样的趋势下，地质数据掌握得越多、越深入，分析越精细，开采的成本和风险就有可能降低。业界也因而出现了“数字油田”（digital oilfield）的新名词，指的就是在探钻过程中广泛收取和分析数据的新型营运模式。企业会在探勘设备和运输管线上大量安装传感器，以实时提取数据，并利用高速通信和数据挖掘的技术，远程监控和随时调



整钻井作业。

业界评估，数字油田的效益如果发挥到极致，可以把产量提高 8%。以原油价格 100 美元的保守估计来看，就算是日产量仅有 1000 桶的小型油田，一年下来也能增加近 4 百万美元的营收。说得夸张点，对石油公司来说，挖数据简直就像挖石油一样有价值！

另一个和一般人生活更接近的例子是，每个月家家户户都得缴纳的电费。传统的机械式电表是一个转动的银色圆盘，用来显示用电度数，但却无法提供详尽的用电信息。大多数人想了解日常用电的唯一信息来源就是每个月的账单。

现在新的以液晶显示器呈现用电量的智慧电表（smart meter），除了可以呈现每一户详细用电量的变化之外，更重要的是通过不同的电价方案，促使用电户自发性降低电能使用，或选择在电费比较便宜的非高峰时段使用洗衣机或洗碗机等高耗电量的家电。这样做，可以帮助电力公司运用电价诱因和缓负载曲线，以降低生产成本。而整合同步向量（synchrophasor）、错误侦测和智慧电表等各种设备运作数据的智慧电网（smart grid）系统，可以形成可视化的消息，提供给调度人员，让他们了解电力设施是否正常运作，一旦察觉到某些设备出现电压不稳或用电量暴增，控制中心也可以迅速进行调度，避免在尖峰时段供电不足，或是在非高峰时段电力闲置的状况。

不过，相较于传统电表每个月只记录一次用电量，以液晶显示器呈现用电量的智慧电表，最少每 15 分钟记录一次。这相当于每个月有 3000 笔记录。以家庭户数约 100 万的台北市为例，光是每个月的电表数据就从



100万条暴增为30亿条。

数十亿条的用户数据、数千个变电站和中转单位、电力传输的复杂计算法则和预报性气候建模软件，这一大片包含了来自网络和电力系统中大量设备与传感器的数据，代表的是能源企业在发展智慧电网的同时，首先要面对的是数据资料蔓延带来的挑战。

另一方面，传统电网中发电端的产电量通常是高度可预测的，所以电力公司可以很容易地配合用电量的趋势，在配电端建立供电模型。然而，因为大量可再生能源，例如风能和太阳能被投入电力系统，这些因气候而导致产能变化幅度很大的新能源，却反而让发电端出现了新的变量。

相较于传统电网无法预测不稳定的电力来源，智慧电网则可以采用同步相量技术，以每秒30次的频率向控制中心回报电路的电压、电流和频率等数据，以提供发电端实时的变化，调整配电的稳定度。但如此做法，不可避免地增加了大量的数据资料。

对于必须时时刻刻维持正常运作的电力公司来说，要建立完整的智慧电网架构，亟需一套新的工具来管理两端的电力平衡和暴增数千倍以上的数据资料。而大数据分析，正是能源企业口中所说的“复杂事件处理引擎”（complex event-processing engines），也是智慧电网能否成为国家级能源政策的关键。

根据中国台湾地区资策会产业情报研究所（MIC）统计，2010年全球智慧电网市场规模约有900亿美元，预估至2015年时将增长至1900亿美元；其规划与发展上以美国最为领先，欧盟、日本、韩国、中国等



也积极展开相关计划。

为什么美国可以率先在 2009 年宣布斥资 34 亿美元，协助公共事业企业在全美 49 个州建构智慧电网，并补贴 1800 万个家庭装设智慧电表呢？其实这也和美国政府先一步重视大数据分析的应用息息相关。

智慧供电系统

南加州爱迪生电力公司（Southern California Edison，SCE）是全美最大的电力公司之一，服务范围涵盖加州中部、南部和沿海地区，每天供电给 500 万用户，其中包括 30 多万家企业，服务总人口近 1400 万人。

2009 年 9 月，在加州公共事业委员会授权下，SCE 推出"Edison 智慧连接"（SmartConnect）计划，预计在 3 年内协助 500 万用户改装智慧电表。用电户不仅可以运用智慧电表调整用电模式，家中有电动汽车的用户还可以在车库中装设家用充电装置，依据自己的使用需求向 SCE 订购充电模式，如夜间充电或白天充电、以 110 瓦功率慢充或 240 瓦功率快充等。

对于目标是在 2050 年前将个人交通设备全面更换为零排放车辆的加州来说，此项计划预计将协助用户减少 1000 兆瓦（约 100 万度电）左右的用电需求，相当于一个普通发电厂的发电量，而整个加州每年也将减排 36.5 万吨的温室气体和排烟污染物，等于路上少了 7.9 万辆汽车。

未来，这项计划更将结合第三方开发新的消费应用层面，例如帮助消费者进行在线用电管理，或是建立 GPS 和电表之间的联系，这样用户



就可以在回家前 20 分钟发送指令，预先打开家里的空调。

对于电力公司来说，智慧电网除了可以自动区分不同的电流，并收取不同的电费，甚至还能进一步自我修复故障。以往，电力设施出现故障时，工人通常有两个选择：一是毫无头绪地搜索故障的根源所在地，二是等待使用者投诉，然后根据投诉人的位置确定大略的故障发生地。

不过，无论是哪一种方法都耗时甚久，因为传统的一个电网区域就广达方圆 1.3 万平方千米，只要一道闪电伴随着一声雷响划过天空，一根树干应声倒地，压倒了电线，就有可能造成数十万用户停电。

通过智慧电网，以前需要好几个小时才能排除的事故，现在只要 10 秒钟，电网就会通知总部哪些电线受到了影响，并且自动改变送电线路，恢复供电。系统还可以依据电路中断的情况通知总部事故发生的地点，以便尽快派遣维修工人前往修复。在饱受电网故障困扰的地区，缩短电力系统停运的时间可以节约数百万元的成本。

应对电动汽车增多而开始实施的西北太平洋智慧电网示范项目（Pacific Northwest Smart Grid Demonstration Project）也是如此。原本美国政府和企业推动电动汽车就是为了节能减排，但是隶属美国能源部的邦威电力管理局（Bonneville Power Administration，BPA）却发现，傍晚通常是电力负载的高峰，因为这个时间大家下班回家开始做饭，并打开热水器或空调等多项用电设备，但人们又习惯下班后将电动车直接停进车库充电；如此一来用电量就会瞬间飙高，很容易使得区域电网负载过重，可能得盖一座新的电厂来支应这项新需求。

由于盖一座新电厂本身就是一次能源消耗，这样做就完全无法达到推动电动汽车以节能减排的目的，于是他们开始思考，怎么样才能在不盖新电厂的情况下提高电网效率。BPA 和巴特尔公司（Battelle）以及华盛顿大学合作研究发现，如果车主可以改在半夜的非高峰时段替电动汽车充电，就可以消化傍晚用电高峰时 70%的电量负载，而且完全不用增建电厂。

但是，总不能要求车主半夜起床替电动车充电吧！他们的解决办法是，针对美国 5 个州、6 万个用户装设智慧电表。这种智能型电表借由感测、接收的用电量数据，帮助电力公司更有效地分配电力，不仅可以在用电高峰期提醒用户，建议关闭某几项用电量较高的空调、干衣机等家庭电器，并依此给予电费奖励，更可以直接设定电动汽车的充电时间，鼓励用户将之设定在电费优惠的半夜时段。

如此一来，BPA 就不必再随着尖峰爆量而筹盖新电厂，或至少可先在电网内进行适当调度，而把新电厂的投资往后延几年。建立一座新电厂从资本投入、营运到维持，通常需要耗资 6 亿到 10 亿美元的费用，可见省下的成本非常惊人。

驾驭气候数据

丹麦的风力发电机组厂商 Vestas Wind Systems A/S 是全球最大的风力发电机供货商，2009 年全球市场占有率为 12.5%。该公司成立于 1945 年，自 1979 年起投入风力发电系统的研发、制造、销售和维修业务，至今已在



全球装设 4.3 万多台风力机组，目前平均每 3 小时就新架设一台风力机，每年总发电量已超过 9 000 万兆瓦（约 900 亿度电），足以供应数百万家庭使用。

Vestas 认为，到 2020 年时，风力发电将占全球一成的电力供给量，有很大的发展潜力。风和石油、天然气等石化燃料不一样，可再生、又干净，也能大规模商业运转。但风力机的选址和配置却是非常棘手的问题。

风力发电是利用风力带动风车叶片旋转，来促使发电机发电。一般认为，风大的地方就适合装设风力机组，但其实，不只风的强度，机组所在风场（wind farm）地面的粗糙度，甚至建筑物、大树灌木丛都会对风速造成影响，而无法产生最大电量。

还有，“紊流”（turbulence）也很麻烦。紊流也称为乱流，是以不规则和不稳定状态流动的气体。常坐飞机的人应该都有遇到乱流的经验，轻微的话，机体摇动或震荡一下就过了，但严重时可能会造成飞行安全事件。风力机也会遭遇乱流，因为机组得朝迎风方向设置，而风流经叶片后会在背风面呈现不规则流动。强劲紊流会降低风能并增加对机组的磨损，造成零件故障。风力机每座造价至少 330 万美元，预计寿命为 20 年，如果没有控制好紊流而导致机器寿命缩短，损失可是非常惊人。

说穿了，风力机位置的选择直接关系客户的投资回报率和 Vestas 的营收。地点选错了，发电量不如预期、机器寿命变短，会减损电厂的投资回报率，导致 Vestas 流失客户。再者，由于机器故障率高，Vestas 也得



负担更高的保固成本。

一直以来，Vestas 靠参考内部的“风力图书馆”（wind library）为风力机选址。这个图书馆其实是个大数据库，里面存有 Vestas 在全球 66 个国家已架设机组位置的气流信息，以及来自世界各地共 3.5 万个气象站的气象数据。

统整这些数据后，Vestas 把全球划分成一个个面积 729 平方公里（约等于一个石门水库的大小）、长宽均 27 公里的正方形区块，就好像地球仪上经纬度交错所产生的网格一样，不过每一格显示的是气象信息。工程师可以运用流体力学的计算模型，把范围再缩小到只有 100 平方米，以评估特定地点的风力、风向和温湿度等。

或许你觉得 100 平方米已经很小了，但如果能把每个网格的面积再缩小一点，气候模型的仿真就能更精准；但前提是要有更多的数据。Vestas 因此计划收集更多气象数据；他们预估，长期下来风力图书馆的馆藏可能会增加到 24 PB，如果都用 DVD 储存的话，这些 DVD 迭起来可以从地球到月球来回 24 趟。

数据变多，其实加大储存设备的容量即可，真正的难题在于如何对更大量的数据做更细致的分析，而且，还得大幅加快分析的速度，才能保持竞争优势。于是，Vestas 和 IBM 合作部署一个新的超级计算机平台，运用专门处理大量结构和非结构数据的分析技术，让工程师可以更准确、更快速预测特定区域的气候模式，以找出发电量最高的位置。

为了决定装机地点，Vestas 过去会分析 178 个变量，包括温度、气压、湿度、降雨量、风向和风力等，以评估当地是否具备适当的气候条件。



所需的分析时间平均要 3 个星期。现在又加入卫星空照图、过去 10 年气候数据、全球森林砍伐指数及地理空间、月亮、潮汐变化等参数，分析的变量暴增为好几百个，但只需要 3 天就能跑出结果。

除了速度变快以外，分析的准确度也提高了。本来长宽各 27 公里大小的网格，已经缩小到长宽都只有 3 公里而已，面积几乎缩小 90%；利用流体力学模型演算后，又可锁定更小的范围，提高分析模拟的准确度和效率。由于 Vestas 可加快风力机的选址和装设作业，机组能够提前一个月上线，让客户提早回收投资。

还有，Vestas 还可检查每一台已装设风力机的运作数据，再与大量的气候信息交叉比对，预测出每个位置的潜在风险，在机器出问题前提早应对，因此能保障客户的投资，也为自己节省产品保固的成本。

调节水坝发电

贵州乌江水电开发有限责任公司是中国华电集团公司的持股子公司。华电集团是中国五大国有发电公司之一，由国务院直接管理。贵州乌江水电开发有限责任公司是中国第一家流域水电开发公司，华电集团持有该公司 51%的股份，另外 49%为贵州省政府所有。

中国是世界水能资源总量最丰沛的国家，且有 75%的水能资源集中在西南部的云南、贵州、四川、重庆和西藏等地，所以素有“世界水电在中国，中国水电在西南”的说法。乌江是贵州省第一大河，也是世界第三大河长江的最大支流。乌江沿岸共设有 9 个水坝，其中有 7 个属贵



州乌江水电开发公司所拥有和管理。

该公司在这7个水坝和其相应的水电站设置传感器，以监控水位、水压、电压、电量和水坝安全机组的状态，但信息没有汇总和分析，所以水电站之间彼此脱节，只是各管各的。

水力发电是利用水坝和发电站之间的水位落差推动发电机，而得到电力。这些水坝设在乌江上游到下游不同区域，每一个都会放水发电。上面的水坝放水后，使上游发电站能够发电，水也会流入下面的水坝。万一上游的水坝一下子放太多水，可能会超过下游发电设备的处理能力而造成浪费，甚至导致下游水位过高而发生溢流，因而破坏沿岸的生态环境；但若水放得不够多，下游发电站则无法产生足够的电量。因此，7个水电站彼此之间其实是互相连动而无法分开运作的。

为了提高总发电量，并降低水坝对环境的伤害，贵州乌江水电公司建立一个整合分析平台，集中管理和调节所有水坝和水电站的营运。通过实时分析传感器所采集的信息，他们可以计算出附近的下游发电站需要多少的放水量，还能针对每个水坝的水文、水位和发电量进行逐年或更长周期的对比，如参考过去几年的降雨量，以更准确地预测所需放水量。

借由精密的计算和预测，现在水坝只会放出发电站所能处理的水量，并同时考虑放水对下游沿岸生态和居民用水的影响。因为贵州近几年饱受旱灾所苦，2011年时东北部地区的旱情甚至持续半年之久，所以，这样的分析能力格外重要，才能确保每次放水都能充分运用，既避免不必要的浪费，也提高总发电量。



延伸应用

能源产业正遭逢前所未有的变革，包括整建智慧电网、增加再生能源、提高发电效率等，挑战着企业处理大数据的能力。上述案例中，SCE 通过推动用户安装智慧电表促成自发性节电，因而省下一个发电厂的发电量。SCE 服务 500 万用户，和台湾电力公司的用户规模差不多，再加上中国台湾省平均每人每年用电量高居亚洲第一，电力公司若采用智慧电表，应该也能创造可观的节能效果。

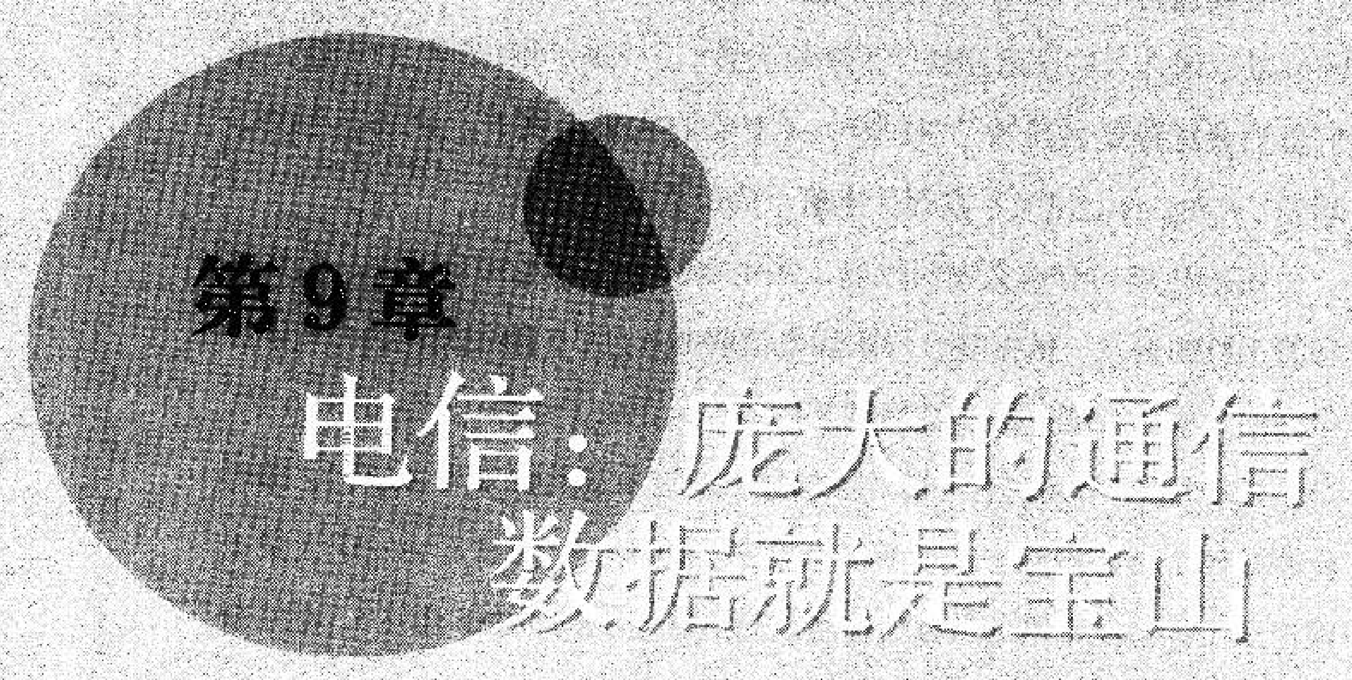
另外，台湾地区近年来积极发展再生能源产业，省内也已建立不少风力和水力发电的设施，如果能在现有基础上加强数据分析的能力，应有助于提升再生能源对整体发电量的贡献度。譬如，沿海地带虽有强劲的气流和季风吹袭，但由于气候潮湿易造成机件腐蚀，可能影响机组寿命。Vestas 参考全球气象数据和已架设机组运作信息来选址的方式，或可作为参考。

除了这几家公司的经验以外，能源产业还能从以下几个方面运用大数据分析。

- **整合移动数据：**移动设备日益普及，消费者也会用手机或平板装置支付水电费、查询账单信息，或在社交网站上询问附近邻里的停电信息。电力公司可分析消费者在网站上的活动模式和意见，进而发展出更符合用户需求的服务，并推动节能减排措施。

- **链接恒温控制器数据**：在寒带国家，恒温控制器是家中很普遍的设备。智能型的恒温器就像智慧电表一样，可以频繁记录和传输用户家中因调节温度所耗用的电量，每一台一个月会产生好几万条记录，善加利用的话，也有助于电力公司调节用电，并鼓励用户改变用电习惯。
- **研究电动汽车车主充电习惯**：在电动汽车较为普及的地区，车主若都在同样的时段充电，用电量会瞬间飙高，容易导致区域电网负载过重。通过追踪和分析车主的充电习惯，电力公司可了解哪些时段供电比较吃紧，并鼓励用户在非高峰时段充电。





第9章

电信：庞大的通信数据就是宝山



星期五晚上 7 点，Kelly 和姊妹淘去最喜欢的意大利餐厅吃饭，慰劳自己一周的辛劳。

餐厅人多，服务生忙不过来，在等候带位的时间里，两人拿起手机玩起互拍照片。几分钟后，Kelly 选了张满意的照片，连上中华电信的移动 3G 网络，正打算更新 Facebook 上的大头照时，看到新消息却忍不住哀嚎："惨了！我老板两分钟前才在这家餐厅打卡！"

听到这句话，Kelly 的朋友知道，她老板是在 Facebook 上标示他所在的位置，现在人应该就在餐厅里享用大餐。但若时间倒转回 5 年前，她们恐怕会觉得莫名其妙，"打卡"，不是在打卡钟上刷卡，记录上下班时间吗？Kelly 的上司又不在餐厅上班，为什么要打卡？

短短几年间有这么大的改变，最主要的原因就是移动设备，尤其是智能手机的普及。依据联合国国际电信联盟（ITU）的预测，2011 年时，全球使用中的智能手机已经超过 5 亿部，到 2015 年时则会增加到 20 亿，等于几乎每 3 个人就有一部。通信技术的进步虽然带来许多便利，随时就能拍照上传到网络上或分享实时动态，但由于移动设备会发送大量上网记录（像智能手机的数据传输量就高达普通手机的 20 几倍），电信企业的系统数据流量也呈现空前增长。

其实，电信公司的营运过程中本来就会产生很多数据，例如，每通电话都会产生一个通话记录（call detail record，CDR），其中涵盖的数据有拨出端的电话号码、接听端的电话号码、通话开始、结束和持续时间等。不过现在因为移动设备普及，3G 上网用户大量涌现，导致 CDR（包括语音和数据传输）数据量更是以等比倍数暴增。

以中国一家数千万用户的电信公司为例，其CDR数据每天新增约5～8 TB，如果以一张容量700MB的CD-R光盘存放，等于每天增加7500到12000张光盘，从上到下一张张迭起来更是有9～15米高。其中，语音数据超过100 GB、短信数据100～200 GB、GPRS数据48 GB，3G数据约300 GB。这些数据都必须完整地保存和处理，电信公司才能准确地计费，但这已经对系统造成相当大的负担。

更何况未来还会出现比3G数据性能更高的无线宽带网络，如LTE（Long-term Evolution，3G过渡到4G网络的一种技术标准）和4G等，让移动用户可以传送更大量、种类更多的数据；每种数据的传输不仅更实时，还有个别不同的服务需求，数据“大、杂、快、疑”的挑战势必又将加剧。国际组织UMTS Forum便预计，2020年全球每年的无线网络数据流量将达到127 EB，足足比2010年时高出33倍。

浩瀚庞杂的数据虽然令电信企业疲于应对，却是藏有用户习惯和营运情报的宝山。营销学上有句老话“若失去一个客户，之后想要把客户找回来，就必须得花费6倍的成本”。对当今的电信企业来说，这个论点的真实性恐怕更甚以往。因为电信公司的用户流失率（churn rate）每升高1%，利润就可能暴跌几百万美元；如果失去的是每用户平均收入（Average Revenue Per User，ARPU）较高的用户，损失又更为惊人。所以，企业无不绞尽脑汁留住顾客，为此，也已经开始把眼光放在大数据分析的潜在效益上了。

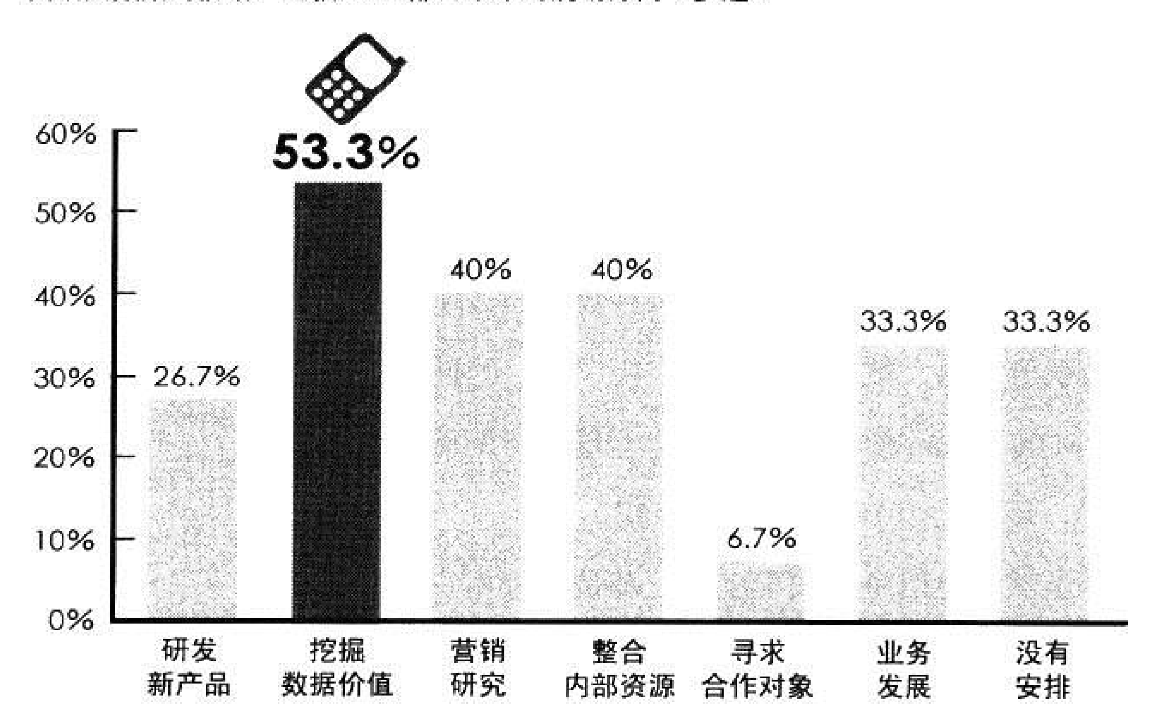
在一份台湾企业IT部门主管（CIO）的调查中，研究人员发现，电信业的客户几乎都是一般消费者，所以挖掘数据价值特别重要。因此，



电信公司最希望能从大数据中挖出各种有价值的信息，以辅助营销策略的规划，进而达到提升企业竞争力的目的（见图 9-1）。

图 9-1　电信业已将目光转向庞大的数据

面临大数据的挑战，电信业 IT 部门未来的规划方向（多选）



数据来源：CIO IT 经理人杂志与 CIO 协进会，《2012 CIO IT 决策者关键调查》

现在消费者越来越精打细算，又很容易就能携款投向对手怀抱。对电信企业来说，客源就是财源，大数据分析则是那把开启财源的密钥。以一家名列《财富》1000 强（Fortune 1000）的电信公司为例，只要能把现有可分析数据的比例拉高 10%，就可能增加 96 亿美元的营收（见图 9-2）。如果把这笔钱再投资在扩充营运上，足足可多雇用近 11.5 万个营运分析人员，或新建 6 万多个基站，这都有助于改善服务质量，进一

步巩固市场竞争利基。

图 9-2 数据分析与电信公司的效益

多分析 10% 的数据，大型电信公司可增加 96 亿美元的营收

电信公司		新就业机会		其他潜在价值
96 亿美元	=	114 825 作业研究员	或	64 000 座基站

数据来源：Sybase

提高客户续约率

中国联通是中国第二大电信商，也是中国唯一一家同时在纽约、香港和上海 3 地上市的电信公司。该公司的 3G 网络广泛覆盖县级以上城市，并且拥有全球规模最大的 WCDMA 网络。2010 年底时，这家公司的用户数已突破 3.1 亿，其中手机用户便超过 1.67 亿；无论在市值和用户规模上，中国联通都名列全世界电信产业的前列。

尽管拥有傲人的记录，服务的又是中国这个全球最大的电信市场，中国联通近几年的压力却有增无减。越来越多的固网通信用户退租而转向移动和无线通信，加上中国每个月有 7000 多万个消费者首次申办手机，造成电信市场区块的变动加剧，连大公司也陷入艰苦的市场保卫战。

中国联通的重庆分公司就面临营收疲软和用户 ARPU 下滑的困境。究竟客户为什么离开？哪些客群最容易流失？什么样的产品可以留住客户？客户想要哪些新服务？中国联通虽然一头雾水，但心里却很明白，答案就在大量的用户数据中！

中国联通决定和 IBM 合作，发展出一套客户流失率分析和营销管理的平台。一般来说，电信用户在退租前会有些迹象，如果能够提早掌握到蛛丝马迹，就有机会留住客人。因此，这套平台广泛搜集各业务部门的数据，归纳出许多会导致客户离开的因素，再比对每位客户过去一段时间内与公司间互动的模式和使用行为，以预测下一个月会退租的几率。

一直以来，中国联通只能粗略推算每个月的客户流失率，而且也无法判断哪些客群会流失最多用户，所以很难锁定特定客户群的需求来加强服务。但现在通过大量分析 CDR 数据，已经把预测流失率的准确性提高 5 倍以上，营销人员因此能针对“危险性”最高的用户，依据他们个别的使用习惯提供更符合需求的方案。

例如，若一个手机用户的使用习惯是以短信为主，但过去 3 个月的短信发送次数却逐月减少，就可能是有心“投向敌营”的高危险群。等营销人员进一步了解后，若发现他采用的是每条短信都要计费的方案，并不适合他的使用习惯，便可建议客户选择短信包月或网内短信免费的优惠方案，以吸引他留下来。

借由精准的预测和营销机制，中国联通重庆分公司的客户续约率已经提高 34%。为了进一步拉高收益，营销人员还针对 ARPU 较低的客群设计更个性化的增值服务和优惠方案，成功地把公司整体的 APRU 提高

10%、短信用户增加13.5%，2G和3G网络的用户数更是跃增6倍。

增加资产使用率

Bharti Infratel（巴提电信基础设施公司）是印度电信龙头 Bharti Airtel（巴提电信）独资成立的静态电信基础设施供货商。电信的静态基础设施，泛指支持电信网络营运的各项设施，包括实体的无线通信铁塔、机房、空调、发电机和火灾警报器等。除了静态基础设施外，还有动态基础设施；这些则是实际电信网络内的设备，如基本信号收发系统、基站、微波设备和传输设备等。

以往电信公司习惯自行经营电信塔，但这么做的费用非常高。研究显示，大型电信商的营运支出中，有六到七成是花在静态基础设施的架设和维修上。因此，国际上已经有一些大型电信商把电信塔和相关设施转入新成立的子公司或合资公司，再把设备资源以租赁的方式提供给同行使用。这种租赁业务在印度特别普遍，因为印度一些偏远地区经济发展较为落后，通信固网的铺设率很低，但移动通信不用挨家挨户铺线，所以需求增长非常快。为了快速扩展移动网络的覆盖率，以及节省架设电信塔的成本，已经有超过八成的电信公司向外租用电信塔。

Bharti Infratel 的主要业务就是提供电信塔的共享服务，就像“包租公”一样，把电信塔租给多家大型电信企业，而这些“房客”所服务的用户规模个个都有好几百万。

如果你以为包租公很好当，那你就错了。在印度约40万座电信塔



中，有超过 3.3 万座为 Bharti 所有。每个电信塔都有必要的电费和营运支出，若房客太少会不敷成本，但房客太多又会影响通信质量；有些房客会积欠房租、有些会浪费电、有些需要修缮设备；而且，最麻烦的是，这类大事小事层出不穷，什么时候发生根本说不准，Bharti 很难掌握资产的状态。

为了更有效率地应用资源，Bharti 和 IBM 合作，建立了一套搭配软硬件的数据分析和实时监控系统。这套系统除了保存 Bharti 内部高达 16 TB 的既有数据，每个月新增的 0.3 TB 业务数据也能从各业务部门实时收集和汇整。

但这只是第一步，数据整合后还得分析。为了要实时掌控旗下电信资产的状况，Bharti 决定针对重要的营运面向订出 34 个关键绩效指标（KPI），包括总营收、营业收入、市场占有率、客户欠款金额、设备正常运作时间和故障时间、每个租户使用的电费，以及“一塔一户”（只有一家电信公司承租的电信塔）的数量。每个 KPI 均有一组默认的参数和目标值，如市占率的参数可能涵盖营收市占率、租户市占率和通信时间市占率等。然后，系统会自动把业务数据比对相应的 KPI，再以图表呈现结果，若有低于默认目标值之处，在画面上还会以红色显著标示。

借由这个方式，Bharti 的管理团队可以很快速地掌握营运现况，也能及时评估现阶段业绩和年度财务计划间的落差，有需要时，还能随时进行不同指标间的交叉分析，以从中发掘改善的机会。

譬如，Bharti 能确实掌握每个地区一塔一户的状况，以及每家电信公司在各区租用了哪些电信塔的数据，由此针对覆盖率较低的企业进行客

制化促销；假使附近有已经“客满”的电信塔，也可说服其中一些租户搬入“住房率”偏低的电信塔。每个租户使用的电费也能时时监控，以分析租户较少的电信塔是否会拉高整体的电费支出，并拟定改进措施。

另外，客户流失和积欠租金一直是营运上很大的隐忧。现在Bharti可持续追踪客户欠款的金额和拖欠的天数，当这两项指标的数值突然升高，可能代表客户对服务的满意度下滑，而Bharti便能尽早介入和排除问题，降低客户流失的机率，或者，也可针对欠款状况比较严重的客户提供协助，进而减少坏账与提高现金流。

延伸应用

随着电信网络的带宽加大、性能加强，电信企业越来越需要大数据的分析能力。

中国联通利用实时分析CDR数据，预测用户退租意愿并规划营销方案，这是电信业目前比较普遍的数据分析应用。电信公司还可从通信数据中找出用户群体里的“小组核心成员”，如一个家庭里最经常联系和被联系的人，留下这个核心用户，就比较容易守住其他家庭成员。除此以外，电信厂商还可在好几个方面应用大数据分析：

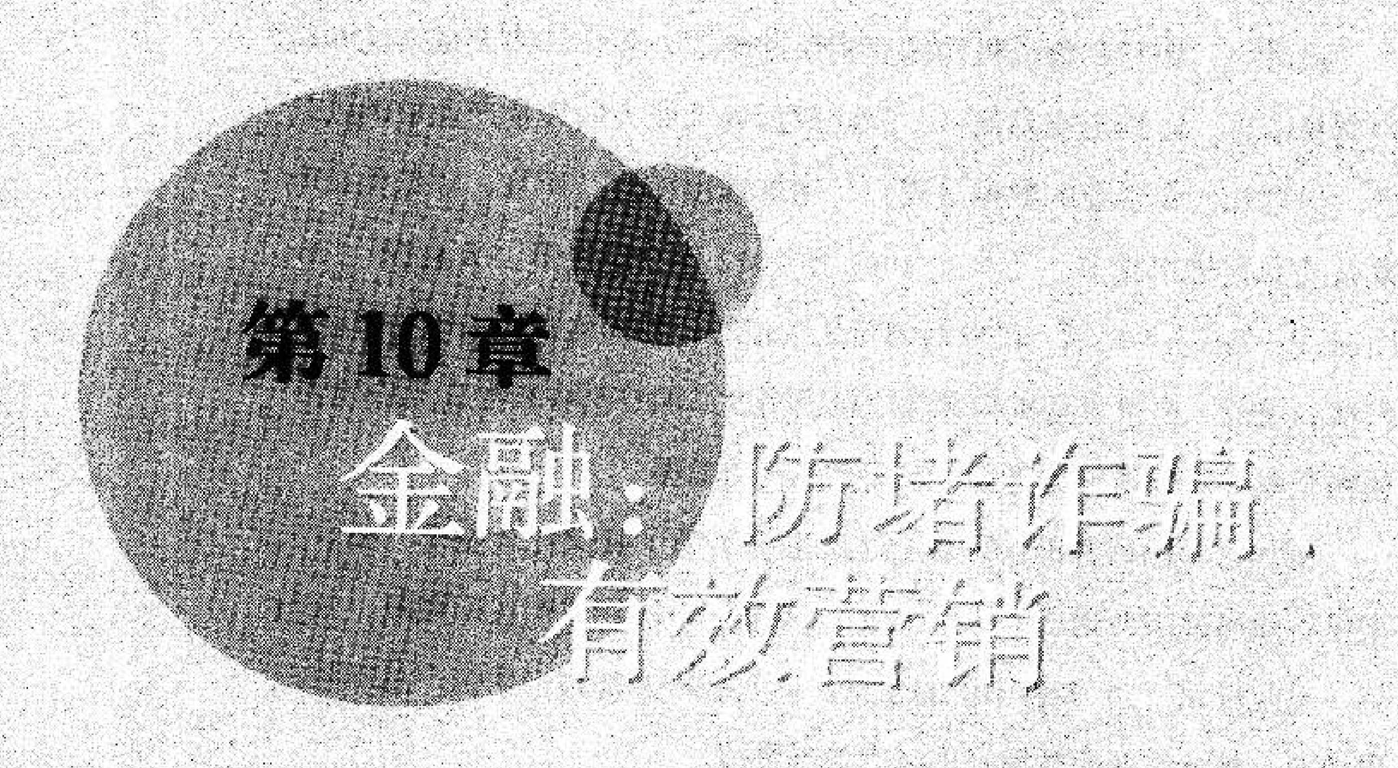
- **改善信号质量**：手机用户在移动中通话时，途中会有不同的基站接收和处理信号，万一某些基站处理能力不够，或恰巧经过信号死角，就会出现断线或通话断断续续的状况。电信公司可分析基



站收讯的原始数据，以改善通信网路中信号较弱的环节。

- **异业结盟，交叉销售**：运用手机发送的GPS位置分析用户的通勤模式，提供合作厂商的促销消息，例如，针对常在中午时间经过某商业区的用户，以电子邮件或短信传送该区域内餐厅的午餐优惠活动消息。
- **过滤垃圾和诈骗消息**：采用流计算分析大量短信，实时拦截垃圾短信和诈骗消息，以保护用户权益。





第10章 金融：防堵诈骗 有效营销



俗话说，“时间就是金钱”，对金融服务产业而言，不仅时间是金钱，数据也是。例如，一般银行在放款前，会先调查贷款人的信用状况、职业和收入，再决定是否贷款给对方；核准信用卡前，也会依照送来的申请书内容和数据，通过各种渠道调查申请人是否具备付款能力，再判断是否发卡。会如此费心，就是因为每一笔数据都可能影响公司的获利。

换言之，银行和投资机构需要明快准确地处理和解读大量市场和客户交易信息，以锁定交易商机，因此数据处理能力向来是产业中的佼佼者。但这几年金融业的数据量变大、增加速度变快，而且，非结构化和复杂的数据越来越多。尤其在 2007 到 2012 年全球经济“大衰退”（the Great Recession）之后，民众对银行失去信心，各国监管机构纷纷祭出更严格的规范，要求金融机构增加信息透明度、加强稽核和提高风险控管，使得企业必须处理种类更多、来源更多元的数据。

例如，美国联邦准备理事会（联准会）为了确保资金安全，要求大型银行须通过压力测试（stress test），才能配发股利或购买库藏股，银行每一季就得填 120 多种不同的表格，以向联准会提报当季资产总和和应用状况，都让企业的数据管理能力备受挑战。

但对当今的金融机构来说，大数据的分析是未来金脉所在。举例来说，流计算可同时跨越多个市场和国家、以近乎零延迟的速度分析海量的交易与市场数据，并在百万分之一秒内迅速计算出结果，让投资银行通过差价交易和业务风险分析转亏为盈。或者，算法交易（algorithm trading）把交易条件写进计算机程序，由超级计算机监测市场信息，一出

现预定状况即由计算机自动下单买卖。这种高频率的交易模式平均每秒可过滤 1270 万条期货买卖消息，并在 0.0001 秒内根据设定为客户提供交易建议。

另外，大数据分析也可协助金融机构减少不必要的损失。还是回到信用卡的例子，信用卡欺诈一直是银行的心头大患，业界数据显示，盗刷和冒用等欺诈行为每年至少造成发卡机构和持卡人高达 5 亿美元的损失。在中国台湾，由于“金管会银行局”规定，持卡人只要发现信用卡有不是自己消费的款项，无需举证，由银行负担所有的损失，所以，为了避免损失，发卡机构努力加强信用卡欺诈的侦测机制。

然而，传统的机制是使用某些模板和预测模型来判断欺诈的模式，只能探查部分的数据，不够实时，也无法把范围缩小到个别事物和个人的层级。问题是，现在诈骗手法越来越高明，常常每几个小时、几天或几星期就会出现新的周期性欺诈模式，如果不能及时获得可识别或支持新欺诈检测的数据，那么在银行发现这些新模式前恐怕已经发生了一些损失。运用大数据分析，信用卡公司可以用更快的速度分析更多的数据，以实时侦测和制止诈骗行为。

除了信用卡盗刷以外，层出不穷的诈保、在线交易欺诈和人头账户等诈骗行为，也常造成重大损失。然而，过去由于实时处理大量数据的技术不够成熟，分析的成本非常高，使得多数企业仅运用不到 5%的可用数据来检测这些非法的犯罪活动。随着大数据技术的发展，现在金融机构已可用更实惠的方式分析剩下的 95%数据。

也因为这个缘故，大数据分析能力极有可能成为决定企业竞争力的



分水岭。试想，一家只能分析持卡人 5%事务数据的公司，和另一家能够多分析 20%的数据的公司相比，哪一家客户受到比较完善的保护、会有比较高的满意度？更何况在欺诈侦测外，不少金融机构已经运用大数据分析了解客户行为、改善投资组合和提供个性化营销等，一场数据淘金战已经开跑，落后者恐怕得付出惨痛的代价。

加快理赔速度

Infinity Property and Casualty Corporation（IPCC）是一家汽车保险商，总部设于美国亚拉巴马州，2003 年于美国纳斯达克挂牌上市。IPCC 是全美前几大的“标准”保险公司，在该公司的车险业务中，高达 8 成属于非标准车险。所谓非标准车险，指的是车主因属于高风险族群，如有肇事记录、年龄过高或投保车种特殊等，不符合“标准”车险申请资格而购买的保险产品。

尽管已经坐稳非标准保险市场，IPCC 还想进军标准保险服务领域，但这么做等于得和许多老字号的保险公司正面冲突，对一家 2002 年才成立、来自南方小镇的公司来说，这是一场不小的硬仗。他们知道，自己一定要有过人的能力，才有机会与全国性的大公司一较高下。

因为这个原因，再加上最近越来越多理赔审核人员向公司反映，诈领保险金的案件似乎有增加的趋势，所以，IPCC 决定发展世界级的保险理赔处理能力，运用一套预测分析机制加强诈保侦测，提升理赔的速度、效率和准确度，以留住老客户和吸引新客户。

IPCC 原来的理赔审核机制高度依赖人为的判断和处理时间，审核人员得仔细留意申请案件是否有诈保迹象，包括保户是否在提高保额不久后马上发生事故，或对理赔流程过于熟悉，若发现可疑案件还得转给其他部门进一步评估，而导致理赔流程拉得很长，影响保户满意度。

在新的理赔系统中，IPCC 仿效信用审核评分的方法，也建立起一套专门评估理赔申请案件“诈保率”的评分机制，一旦发现可能的问题案件，系统会按照事先设定的业务规则，把案件转给负责调查的人员。

例如，在保户发生意外的当下，保险业务会先抵达事故现场收集数据，之后系统按照业务规则评估这个理赔申请的诈保率，假使诈保率超过默认值，申请书就会自动转到理赔调查员的手里。这些调查员的工作好像警探一样，除了要搜查保户数据之外，有时还得出动查访和跟监埋伏，时间拖得越长证据就越不容易取得。以前可疑的案件往往需要一到两个月才能送交调查，现在则缩短为一到两天，因此，这已经让 IPCC 把阻止诈保的成功率从 50%提高到 88%。

为了加快理赔处理速度，IPCC 也从收到保户通报事故的第一时间着手，运用演算模型，在事件发生当下就把理赔申请按不同处理和评估需求分门别类，让有问题的案件可以尽早被调查，而且，不需要调查的案件也可立刻获得给付。通过这样的方式，该公司在第一时间就能排除25%需要后续调查的案件，省下不少案件往来的时间和费用。

另外，IPCC 还采用文本挖掘（text mining）技术，分析警方对交通事故的调查报告、伤者医疗记录和其他文件中的内容，检查描述上有何矛盾或可疑之处，以找到诈保的蛛丝马迹。例如，医疗报告中若出现“药



物”等关键词时，可能和事故中驾驶人的精神状态有关，系统便会启动警示提醒审核人员注意，大幅提高理赔的审理速度和准确度。

同理心营销

做营销的人都知道，朋友推荐比广告有用多了。到底多有用？AC 尼尔森的研究显示，同样的一则营销消息放在 Facebook 上，因为朋友的推荐，网友的点击率会增加 16%，可见同侪之间的影响力有多大。

中国一家国有银行就想把这种同侪间彼此认同的拉力转为品牌的吸力。第一步先从网络银行开始，在 IBM 中国研究院的协助下，发展出一套业界首创（first of a kind，FOAK）的方案，按照网银客户亲朋好友的投资动态来提供产品建议，以帮助他们找到更多投资机会。

IBM 研究人员说这项技术运用的是人类的“社交同理心”，但要激起客户的同理心，前提是得先了解他们的社交模式。因此，系统先从银行各个经销渠道收集客户的个人身份（如年龄、性别和婚姻状态）和事务数据（如存款和投资金额），经过清理和汇整后进行深入的分析比对，找出客群中有哪些人属于同样的社交圈，譬如是不是互为同事或同学，以及在不同的圈圈中各扮演什么样的角色，如专家或先驱者等。描绘出客户群体之间的关系后，这套方案会再分析客户近期的购买倾向，以及已购买产品的绩效，以辅助营销。

你可能正在想，万一有人就是没什么朋友，根本组不成社交圈怎么办？这一点完全不成问题，因为一般人不仅会听亲朋好友的建议，如果

和自己条件相仿的人有成功的经验，我们也会想要“见贤思齐”，所以，同一个圈圈里的人不见得彼此认识，而可能是经济条件、年龄和工作相仿，或事务数据有相同结构的人。

举例来说，小王是工程师，虽然待遇不错，但因为工作太忙而没有时间，也不懂理财，所以只把钱放在银行存款。有一天，他登入银行网银的这套系统，系统显示跟他收入水平和背景相近的人最近的资产配置状态和投资报酬。看过分析后，他发现，懂得投资的人多半只把一部分的钱存起来，剩下的则投资在股票、债券和共同基金等报酬率较高的产品上。小王开始思考，该不该也调整自己的理财方法。他在网上找到理财专员，经过进一步的讨论和当面说明后，他也决定购买共同基金。

借由这样的技术，这家银行走出和以往不同的营销模式。以前银行虽然也有客户分层的机制，但仅能依照客户交易额做粗略的划分，如存款超过 50 万人民币者为 VIP 客户。这样的分法不够细致，现在则更精细地区隔出不同背景、环境、社经条件的客户层，并且避免让客户有被“强迫推销”的感觉，改为提供同侪理财成效的相关信息，激发客户的好奇心，进而鼓励客户购买更多金融产品，加强客户对品牌的认同度。

选择分行地点

开店做生意的人都知道，选择店面最重要的是“地点、地点、地点！”地点选得好，人潮容易聚集，人潮多了，钱潮自然滚滚来。如果只在一个地方做做小生意，选个好地点应该不算太难，但如果你得在中国这个



幅员广阔的地方经营上万家店面，问题就棘手得多。

这就是中国一家国营银行面临的难题。这家公司成立已经超过20年了，在全中国有一万多家分行，名声非常响亮，但由于中国都市化快速发展，各地窜起许多新兴都市和卫星城市，大型银行也得快速拓点卡位。

该在哪里开分行？总不能跟着麦当劳走吧！在理想状态下，银行分行所在位置和所提供的服务必须符合当地经济发展和客户对金融服务的需求，才能创造最佳的业绩。但中国那么大，都市众多，单单以江苏古城苏州市这个“二级”城市为例，就足足有30个台北市那么大，人口超过1000万。撇开北京、上海和广州等一线城市不说，和苏州同属二级的城市还有200多个，另外还有近400个三级城市，每个市场的需求都不一样，因此，分行据点和服务内容的规划真的是门大学问。

虽然各地市场特性不同，但这家银行认为应该可以采用系统化的方法选择分行地点。2006年时，他们找上IBM合作，设计出一套分行网络优化的系统，以锁定最适合开设分行的地点，依此支持新据点的开拓，或辅助既有分行的搬迁决策。

这套系统采用营运研究和管理科学的技术，发展出一套评估市场商机和分行业绩的量化机制，而能依据一地的地理和人口数据预测当地需要哪些类型的金融服务，以及可能为银行带来多少业绩，再从中选出高商机的市场区块和适当的地点。

根据Forrester Research的研究，43%的人和30%的小公司会依照地利之便选择往来的银行。不过，地点虽然重要，但建筑物的租金、设备的费用、人事的成本等等也会影响到一个分行能不能赚钱。因此，这套

系统还考虑银行实际的管理要求，如最低投资回报率、人事成本规划等，把这些信息和潜在据点的条件配对分析，进而决定最佳设行地点，以及业务和人员职能的组合。

2009年起，这家银行开始推广这套分析机制，到2011年中，大约已经完成100个城市的据点调整或设点的工作。该公司估计，通过选择最佳的营运地点，像苏州这种规模的城市，客户存款额就可提高700万元。

苏州只是一个例子，如果每个都市内的业绩都能创造类似成效的话，总体效益非常惊人。此外，有了这样一套以量化方法评估设点效益的机制，这家银行可更快速和精细地检讨设点的成本和潜在商机，一来可以避免因选错地点而导致分行业绩不佳，二来也能加快拓点速度，在开拓新市场时比对手更具优势。

节省营销成本

"先生您好，您是本行的优质客户，我们现在推出超优惠的信用贷款，不知道您有资金需求吗？"很多人都接到过银行这类的营销电话，有时候，同一家银行甚至会打好几次电话推销不同的产品。

银行电话营销如此泛滥，当然和这几年市场竞争加剧，企业急着抢客户有关，但最令消费者不堪其扰的是，营销消息太密集、促销的产品又引不起兴趣，感觉银行根本只是在"乱枪打鸟"。

从一家中型的省级商业银行了解到：客户不想被广告轰炸。如果不能精确地锁定营销活动的对象，不仅会浪费营销资源，更可怕的是，还



可能会气走客户。过去，该公司的营销人员只凭经验推断促销活动该锁定哪些顾客（如某张信用卡年费较高，所以适合年收入较高的白领工作者使用）。但是，一家银行有几十种产品，还可搭配组成组合型方案（如房贷搭配火险，或存款户专属的理财型信贷等），人的经验和判断毕竟有其局限，这种方法不够有效率，也很难评估营销成果。

于是，这家银行导入一个数据仓库和大数据分析的平台。这个平台会从银行各业务系统和销售渠道收集数据，把管理存放款和汇兑等关键业务的核心系统、财务管理、信用卡、电话服务中心等系统的数据都整合起来。因此，营销人员促销时可以统一筛选目标客户，不必再像以前一样追着各部门要顾客名单，而且，更重要的是，通过分析，他们还能精准地执行营销的"P-D-C-A"循环。

"P-D-C-A"循环由美国著名的管理学家戴明（William Edwards Deming）所提倡，4个字母分别代表管理的4个阶段：P是计划（plan），D是执行（do），C是查核（check），A是行动（act）。戴明认为，企业若确实执行这4个步骤，就能持续从错误中学习而创造巨大效益。

这家银行把"P-D-C-A"循环运用在营销管理上。

P&D：在规划营销活动的阶段，营销人员可在系统上设定欲锁定族群的条件，例如，针对新的悠游联名卡，目标客群条件为从未向该行申请过信用卡、每个月存提款超过10次、常用ATM进行小额提款，且风险度为中级以下的20～30岁顾客。

在最短数十秒内，系统就可找到出符合条件的客户，并估算他们可能为银行带入多少营收，还能根据他们过去的营收贡献度和"可营销性"

排出优先级，以促成营销资源最有效的运用。此外，营销人员还可改动设定的条件，比较不同设计的潜在成效，再选择效益最高的方案。

C&A：平台会从每一个销售渠道收集结果，让营销人员随时掌握活动执行的状态及客户的意见，一旦发现成果不如预期，也能及时调整活动内容，加强营销效益。

借由事前分析找出潜在效益最高的方案，该公司预计可将规划和执行促销的成本降低 10%。另外，由于可以更精确划分目标客户群，尽可能让每一通营销电话、每一封 DM 都“正中红心”，客户对营销活动的响应率也可提高 60%，不仅如此，不适合或没有意愿的人会被过滤掉，避免造成客户困扰，进而提高整体客户的满意度。

延伸应用

金融产业的顾客要求越来越高，忠诚度却越来越低。再加上发达国家增长趋缓，各国企业纷纷抢进新兴市场，金融产业现在得面对来自四面八方的竞争对手、日益严格的法规限制，以及更多样化的客户要求，也因此，企业更需要从大数据中找出有用的信息，以加强市场区隔力。

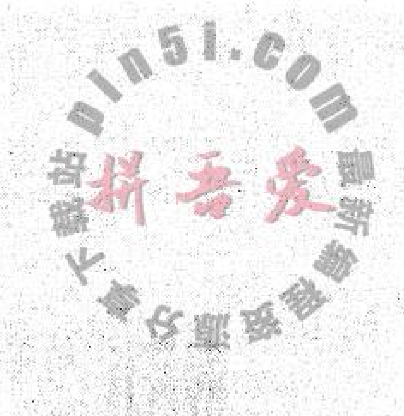
很多公司从改善客户服务下手，上述 4 个案例虽然主要的着眼点不同，但最终的效益或多或少都有助于提高客户服务的质量。以分行设点来说，银行若地点选得好、又能提供民众所需的业务项目，顾客办起事来更便利，就不必再为了同一件事跑好几家银行。

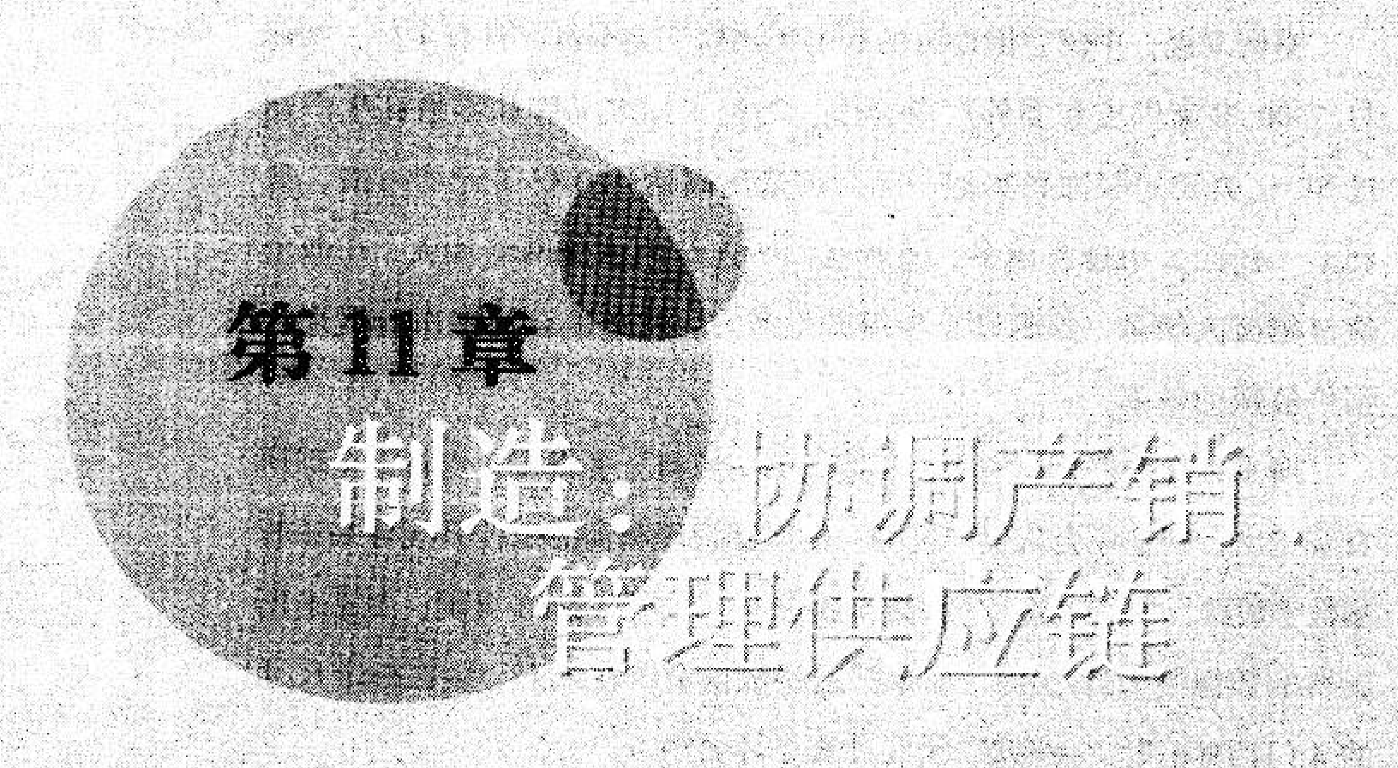
除了本文所介绍的几个案例之外，金融服务产业还可在几个领域使



用大数据分析：

- **跨账户转介分析（referral analysis）**：分析自动转账服务（ACH）的事务数据，包括薪资转账活期存款、房贷自动扣缴等，以挖掘交叉销售的机会，增加银行所管理的总资产规模（AUM）。
- **生命周期营销**：从事务数据中发掘客户生活中的转变，如：更换雇主、婚姻状态改变或购房等，以提供更切合需求的产品和服务。
- **从下到上的（bottom-up）风险分析**：追踪ACH、信用卡和缴付房贷等事务历史记录，分析客户的风险值、及早侦测违约迹象，并视其需求推出促销方案。
- **商誉分析**：投资银行在金融海啸后成为各方箭靶，为了挽回商誉，瑞士银行（UBS）开始大量分析网络、媒体、政府机构、非政府组织、智库、社交网站和博客等第三方信息，并与两万多家现有和潜在客户交叉比对，从中找出可能会影响这些企业商誉的风险事件，如涉及非法砍伐、使用童工或种族歧视等。瑞士银行会把风险过高的公司列入黑名单，并且避免为他们提供服务，或向他们购买服务与产品。





第11章

制造：协调产销，管理供应链



制造业经常是带动一个社会发展转型的火车头，也是经济增长和就业市场的中流砥柱。在成本较低的新兴国家生产力跃进之下，制造业早已成为全球性的产业，但近年来由于资通科技发达和贸易障碍减少，各生产地可针对制造过程中某些环节发展专业能力，厂商为了节省成本，跨国设计、采购、组装、制造、再制、营销和服务的生产网络远比过去扩散和零碎，复杂度更甚以往。

举例来说，中国香港地区的利丰集团旗下完全没有自有工厂，却能与 7500 多家供应和委外厂商合作，为世界知名品牌和零售商生产超过 80 亿美元的服饰和其他产品，创造出高达 120 亿美元的市值。像利丰一样的公司越来越多，随着他们所连结的伙伴网络加大，产业价值链越庞大纷杂，若要进一步提升生产力，就必须设法善用数据来加强价值链的效率。

所幸，在数据分析方面，制造业的“原料”多得很。麦肯锡全球研究院指出，制造业会从生产机械、供应链管理和商品监控系统等来源收集数字数据，本来就是生产和储存数据的“大户”。2010 年时，制造业所新增的数据便将近 2EB，如果把这些数据全印在纸上，装在标准尺寸的 4 门档案柜里，会需要 400 亿个柜子才装得下。

而且，随着企业普遍采用信息系统管理价值链中的活动，并广泛保存来自生产系统以外的数据，包括计算机辅助设计、计算机辅助工程和生产开发管理协作系统等，数据量的增长只会越来越惊人。眼见着大数据排山倒海而来，企业也急着想在被淹没前站稳脚步，研究机构 IDC 在 2012 年中对美国制造商所做的调查显示，认为大数据的管理工具“非常

重要”或“重要”的企业已经超过半数。

制造业可将大数据的管理方案应用于以下几个领域（如图 11-1 所示），以加强价值链（包括研发设计、供应链管理、生产、营销与销售，以及售后服务）现有运作模如下。

图 11-1　大数据的 3i 新世界

	研发和设计	供应链管理	生产	营销与销售	售后服务
❶ 供应链共同的设计和研发数据库	✓				
❷ 客户数据汇整与分析	✓			✓	
❸ 协作平台数据分享	✓			✓	
❹ 供需预测		✓	✓	✓	
❺ 可视化管理生产			✓		
❻ 营运数据分析			✓		
❼ 售后数据分析			✓	✓	✓

数据来源：麦肯锡全球研究院

1. 建立共同的跨部门研发与产品设计数据库，让供应链中的伙伴同步进行产品与程序设计、仿真和实验，并支持内部人员与外部伙伴同步创作。
2. 汇总和分析客户数据，提高服务质量、发掘追加和交叉销售的机会，并落实“价值设计”（design to value），即依照客户喜好和需求研发更具附加价值的产品。
3. 经过虚拟的协同合作平台采集和分享数据，以推动“众包”（crowd sourcing）等创新措施。“众包”一词在 2011 年刚被加入



韦氏大辞典，意思是指企业在社交网站或在线交流论坛上寻求网友的创意和建议，进行产品与服务创新。

4. 建立需求预测和供应规划的机制，改善生产、供应和销售管理。
5. 导入精益生产（lean manufacturing），并运用生产仿真的模型，以可视化工具呈现生产瓶颈，提高生产流程的能见度。
6. 分析传感器所采集到的营运数据，以增加生产力和进行大量客制化。
7. 实时从传感器收集商品售后数据，以及客户对商品的回馈，以辅助售后服务及侦测生产或设计的瑕疵。

上述应用可协助制造企业免除产品研发过程中不必要的重复，及改善生产和组装的流程，以加强整体价值链的生产力，也可让产品更符合消费者的需求，提高产品的价值。除此以外，企业还可开发创新的服务和业务模式，提高竞争力。例如，BMW 的 ConnectedDrive 概念车运用传感器科技，能够依据实时路况为驾驶提供行车方向的建议，当车上传感器侦测到问题时，也会提醒驾驶员维修，并且把车况数据直接传送给维修站。

若把大数据分析应用效益化为实际数字，按照 MGI 的预估，有的公司甚至可将毛利提高 3 成，生产营运成本还可能减半，如图 11-2 所示。

图 11-2 制造业运用大数据分析的效益

价值链环节	大数据应用领域（举例）	成本	营收
产品研发与设计	• 同步工程设计 • 价值设计 • 众包	• +20%～50% 产品开发成本 • +30% 毛利 • -25% 产品开发成本	• -20%～50% 上市时间
供应链管理	• 供需预测	• +2%～3% 毛利	
生产	• 分析传感器所采集的生产线数据 • 生产过程模拟与预测	• -10%～25% 营运成本 • -10%～50% 组装成本	• 最高 +7% 营收 • +2% 营收
售后服务	• 分析传感器采集的数据，改善售后服务	• -10%～40% 维修保固成本	• +10% 年产量

数据来源：麦肯锡全球研究院

产销规划协调

成立于1945年的海泰制果（Haitai Confectionery & Foods Co., Ltd.）是韩国第二大饼干和糖果制造商，总部设于首尔。该公司年营业额近6000万美元，在韩国共设有4家制造厂房，雇用2400多名员工。

面对多变的市场，食品制造商普遍运用销售和营运规划流程来预测供需。海泰制果虽是老牌食品大厂，却苦于无法精确地掌握客户的需求和销售量，影响销售和营运规划的正确性，导致该公司常因存货过多而



损失数百万美元，或因低估需求，生产线应对不及而造成产品质量不稳定。

为了改善销售与营运规划，海泰决定与 IBM 合作导入一套商业智能和分析平台，统一整合和分析海泰的生产、物流和销售的实时和历史数据，并且机动性地追踪每笔订单和每个销售点的供需变动，再分析这些不断更新的大量数据，进而预测哪些地区的商店、在什么时间点会需要哪些商品。

由于这套系统采用了可在线分析处理数据（OLAP）的软件，让海泰从不同角度检查其营运和销售数据，还能剖析各业务部门的销售业绩、客户购买行为和业务趋势有何特定的模式，所以，海泰能够精准地预测销售量，而且几乎立刻就可随市场需求和趋势的变动调整生产规划。

举例来说，韩国虽然属于温带气候，四季分明，但是天气变化很大，偶尔也有热浪袭击，此时冰冻制品就会特别畅销。过去海泰没有分析销售量信息的机制，仅能粗略估算每样商品的销售量，再依据估计排定生产和铺货作业，万一出现天气忽然变暖带动冰冻制品需求上扬时，都会来不及供货而错失商机。现在则不同了，海泰不仅可以迅速发现冰冻制品销量增加的趋势，还能够预测个别销售点和区域会需要哪些种类的冰冻制品，以及需要多少数量。

除了分析之外，高层主管还可使用系统的仪表板功能。因为仪表板采用简单的图表呈现营运的重要面向，所以，包括业绩、生产计划和时间、销售和存货状态，以及销量预测等全都一目了然，让高层主管省下



看一张张报表的麻烦。

借由这个分析平台，海泰可以确实追踪各销售点的销售量，快速掌握市场需求的变动，并运用销售数据进行每日销量的预测，因此，该公司销售预测的准确度已提高7%，而且，存货时间也缩短3天，节省了可观的成本。

全球供应链管理

日本一家照明用具制造商总共生产两万多种照明产品，其中也包括LED灯泡，由于产品阵容庞大，生产线流程相当繁杂，国内外的供货商和业务伙伴也多。

这家公司内部没有集中的供应链管理机制，包括市场需求预测、生产线排程和销售策略的规划等，一直都是由各个业务部门和区域分公司独立执行。当公司仅在日本本地营运时，这种方法还堪用，但随着营运触角不断延伸，他们发现供应链的管理越来越困难，经常出现出货延误的状况。

供应链管理涵盖从供货商到终端消费者之间，包括采买原物料到配送商品的所有过程，每一个阶段都环环相扣。这家公司和很多企业一样，其供应链日益全球化和互相连动，即使是很小的步骤和计算的错误，都会让后果不堪设想，影响整个复杂的供应链网络，造成产销之间出现断层，也冲击公司的经营绩效。

为了解决这个问题，他们打算建立一套全球整合的供应链管理系统，

把采购、生产、销售、库存和供需规划全都集中管理，并借由这个机会加强各生产线和生产基地之间的协调。

他们找上 IBM 公司，由 IBM 建立一套供需管理的系统，统筹管理国内外的生产线。除了整合各分公司的产销和财务等流程外，系统还能够评估会影响供需的各项因素，包括市场需求的季节性波动趋势、气候变化和天气型态的现况，以及非预期性的事件或意外等，以模拟出可能会对供应链带来什么样的影响，再用简单的可视化方式呈现出来，让管理人员快速掌握供应链中较为脆弱的环节，以提早预测可能发生的问题，进而改善某些特定流程。

例如，依据分析，下个月欧洲市场对 LED 灯泡的销量预计将增加25%，散热模块和其他部分关键零部件主要产地在中国，但有地利之便的中国工厂却因连续几个月接到大单，生产线已经排得很满，如果要等中国生产线空出来，不知道会不会延误出货，后续又会不会影响到铺货和仓储的作业？如果到离欧洲市场较近的土耳其生产，关键零部件的几家中国供货商有没有能力在指定的时间内供货，如果没办法的话，是不是应该改向几家小一点的供货商采购？

该公司可以模拟几种不同的情境，包括：在中国生产和备料会比预计供货时间延误几天；如果换到土耳其生产，全部向中国供货商采购物料，或者只向中国供货商采购 30%、40%或 50%，其余比例则发单给越南的供货商等。

通过这种模拟的方式，该公司可事先比较几种情境对供应链的影响，及最终会带来多少销售和财务绩效，以找出合理、潜在效益最高的排列

组合，避免因生产不及而延误出货。该公司已经把准时出货的比例提高到 98%，让销售更有保障，而且，由于事前已评估最佳的生产、销售和库存选项，还能省下各环节不必要的浪费，而强化财务表现。

延伸应用

当鸿海郭台铭董事长都自叹已经从“毛利率 3%～4%”变成“毛利率 1%～2%”，就表示连高科技制造企业都再也不能走削价抢单的回头路。面对成本几近见骨的微利趋势，唯有更积极快速的创新，才能在严酷的竞争中存活，甚至大放异彩。

以大数据分析为核心的创新研发能力，将攸关制造企业如何在日益艰辛的订单争夺战中脱颖而出。研发团队可通过大量的意见回馈，改善测试的周期时间、质量和效率，让企业更快响应市场，抢得先机。就生产制程面来说，实时数据分析能让生产线上的各种蛛丝马迹都纳入观测，并且不断进行实时优化，以减少重复错误所导致的成本与时间耗损。若能将这项效益进一步扩大导入供应链的协同整合作业，实时聚焦来自上中下游的内外部大量数据，提升决策的效率与质量，也有可能进一步提升竞争力。

制造企业还可于以下几方面应用大数据分析。

- **预测分析**：广泛收集生产、销售、研发、存货和运送过程中的数据，预测可能发生的风险事件，以及早防范和改善。



- **生产质量管理：**增加生产活动各环节信息的能见度，并发掘减少浪费、事故和环境污染的新方法。
- **整合生产排程和经营：**运用机动性的任务规划和排程工具，以及先进的模拟技术，找出最恰当的排列组合方式，提高整体的生产效率和生产量。
- **生产能见度和可视化呈现：**信息要让人能够了解和行动，才是有用的信息。现在厂商的价值链散布各地，光是管理厂房就已经够复杂的，更何况还有供货商和经销商等众多伙伴。若能把繁复的营运信息，包括人力、系统、存货、生产线等，以简单的图、表或指标的方式呈现，管理人员就能及时掌握各环节的绩效，并更快速地采取行动。



第12章

娱乐：更深入、更实时的娱乐体验



你可注意到“看电视”变得不一样了？

公司加班害你错过晚上 8 点播出的连续剧吗？不要紧，你可以回家洗完澡后再舒舒服服地用计算机在线收看或下载收看，好在下一集播出前赶上进度。

这就是互连消费（connected consumers）的时代。这时代的消费者有三多：内容多、屏幕多、频道多，他们不再被动“接收”信息，而是通过各种各样的设备，在想要的时间和地点主动“使用”内容。他们所接触的信息和取得内容的渠道远多于以往，在电视、手机、计算机、影音播放器和平板设备间不断游移，找寻需要的和喜欢的内容。

在这样的趋势下，传统印刷和广电媒体的数量变少，影响力也逐渐式微。IBM 商业价值研究院（IBM Business Value，IBV）2011 年访问近 4000 名消费者后指出，中国、日本和美国观众看电视的时间已经大大缩短，而移动影音服务的使用量相对增加。

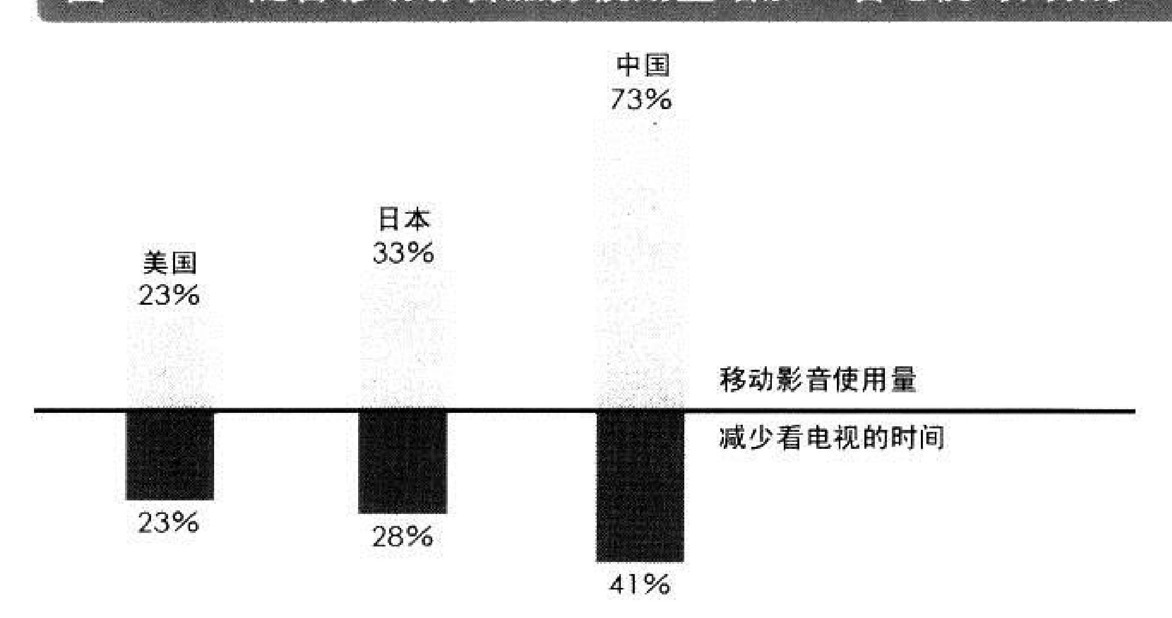
面对一心如此多用的消费者，媒体和娱乐公司该怎样抓住他们的目光？又该如何提供消费者感兴趣的信息？随着视听族群的口味越来越刁钻，争夺消费者时间的平台又越来越多，IBV 指出，企业必须走出过去以提供内容为主的营运模式，发展出独树一格的价值。

要做到这一点，他们得从席卷而来的数据狂潮中找寻商业机会，并愿意运用新发现来驱动决策，这样才有机会从消费者和对手手中抢回主导权，但是，速度要快，否则难保不会重蹈音乐产业的覆辙。

在 2010 年的一篇文章中，CNN Money 将 21 世纪前 10 年称为“音乐失落的 10 年”。Forrester Research 也指出，1999 年的音乐零售总销售

量为 146 亿美元，2009 年时则已萎缩到 66 亿美元。但事实上，很多人比以前更常听音乐，音乐总市值有增无减，问题在于商机已经从传统唱片公司转向其他渠道。

图 12-1 随着移动影音服务使用量增加，看电视时间减少



大数据分析可以帮助媒体或娱乐公司改变营运模式，在产业翻盘前及时响应变动的市场需求。企业可通过整合供应链内各种业务活动，了解业务数据被管理、追踪和利用的方式，控制营运的风险。例如，日本某电视台在网络上针对收视率高的节目搜索关键词，分析观众的意见，研究观众感兴趣的主题和趋势，作为制作新节目的参考。

另一个大数据分析相对普及的领域就是职业运动，这个行业因为高利润、高风险，营运不确定性特别高，所以一些职业球队开始采用先进的分析软件辅助决策。最有名的就是 NBA 的迈阿密热火队（Miami Heat）总教头 Erik Spoelstra，他利用统计分析软件拟定战术和排出上场球员的名



单；俄克拉荷马雷霆队（Oklahoma City Thunder）则靠分析数据选拔新球员。

大数据分析还可用来提升消费者的视听体验。例如，不少电视台已经导入实时分析的软件，在重要运动赛事直播过程中随时分析选手的表现，所以，当某篮球队的控卫站上罚球线，每投出一球，主播就能分析球季到此时此刻为止他的罚球命中率多少，这样的表现在联盟中的排名第几，还差多少就能打平历史纪录等。

另外，为了应对“三多”视听族群跨平台和跨频道的消费习惯，企业还可从大数据中发掘跨平台服务和销售的机会，以增加收入或强化品牌定位，以下温布尔登网球公开赛的做法就是绝佳的案例。

深入实时的观赛体验

温布尔登网球公开赛（Wimbledon Tennis Championships）是目前唯一在草地上举行的网球锦标赛，通常于 6 月或 7 月举办，是每年网球大满贯的第 3 项比赛，排在澳洲公开赛和法国公开赛之后，以及美国公开赛之前。每年在历时两周的赛程中，将近 50 万的球迷涌入球场观战，媒体也从世界各地蜂拥而至报导赛况，而他们的转播会通过 129 个频道，播送到 173 个国家，吸引 7 亿多观众收看。

温布尔登是网球运动中历史最悠久和最具声望的比赛，当然是收视率的保证，只不过现在消费者的选择太多了，不仅其他体育节目想抢观众，连娱乐频道、影音娱乐甚至电玩等其他平台也想来分一杯羹。为了

维持住品牌的光环，并吸引更大量更多元的收视族群，主办单位全英俱乐部（All England Club）希望能让球迷身历其境地体验赛事，而且用更活泼和有趣的方法彼此互动和分享意见。

IBM是温布尔登网球公开赛的官方科技顾问和赞助商，从1990年起就协助全英俱乐部开发和管理信息系统，将温布尔登公开赛打造成最智能的专业网球锦标赛。2011年时，IBM更采用云技术把温布尔登官方网站改头换面，不仅强化网站的各项功能，确保网站可承载高达上亿次的点阅率并平稳运作，还推出深度分析技术提供实时赛事分析。

国际上很多大型运动赛事，如世界杯足球赛和奥运等，主办单位除了得筹办实体的比赛，还要应付庞大的转播及在线浏览需求。为了短短一个月或几周的比赛，主办单位就得大手笔买主机、建机房和架软件，光是采购设备动辄就要上千万美元，但网站的效能可能还不怎么样。

温布尔登网球公开赛期间，大会官方网站（Wimbledon.com）每日造访人次会暴增到1600万，网页浏览量更有4.5亿次，每天网站要处理的信息安全事件最多可能有8万件。采用云端架构，网站可以机动增加或减少计算资源，以弹性应对赛期和非赛期中不同的浏览需求，而不必特别为了比赛而采购新的主机。

IBM提供给温布尔登大会官网的这套云服务架构，也同时支持其他重要的年度运动和娱乐盛事，包括另外3大网球公开赛、百老汇戏剧大奖东尼奖（Tony Awards）颁奖典礼，还有美国高球公开赛（US Open Golf）和高球名人赛（The Masters Tournament）。由于这些活动



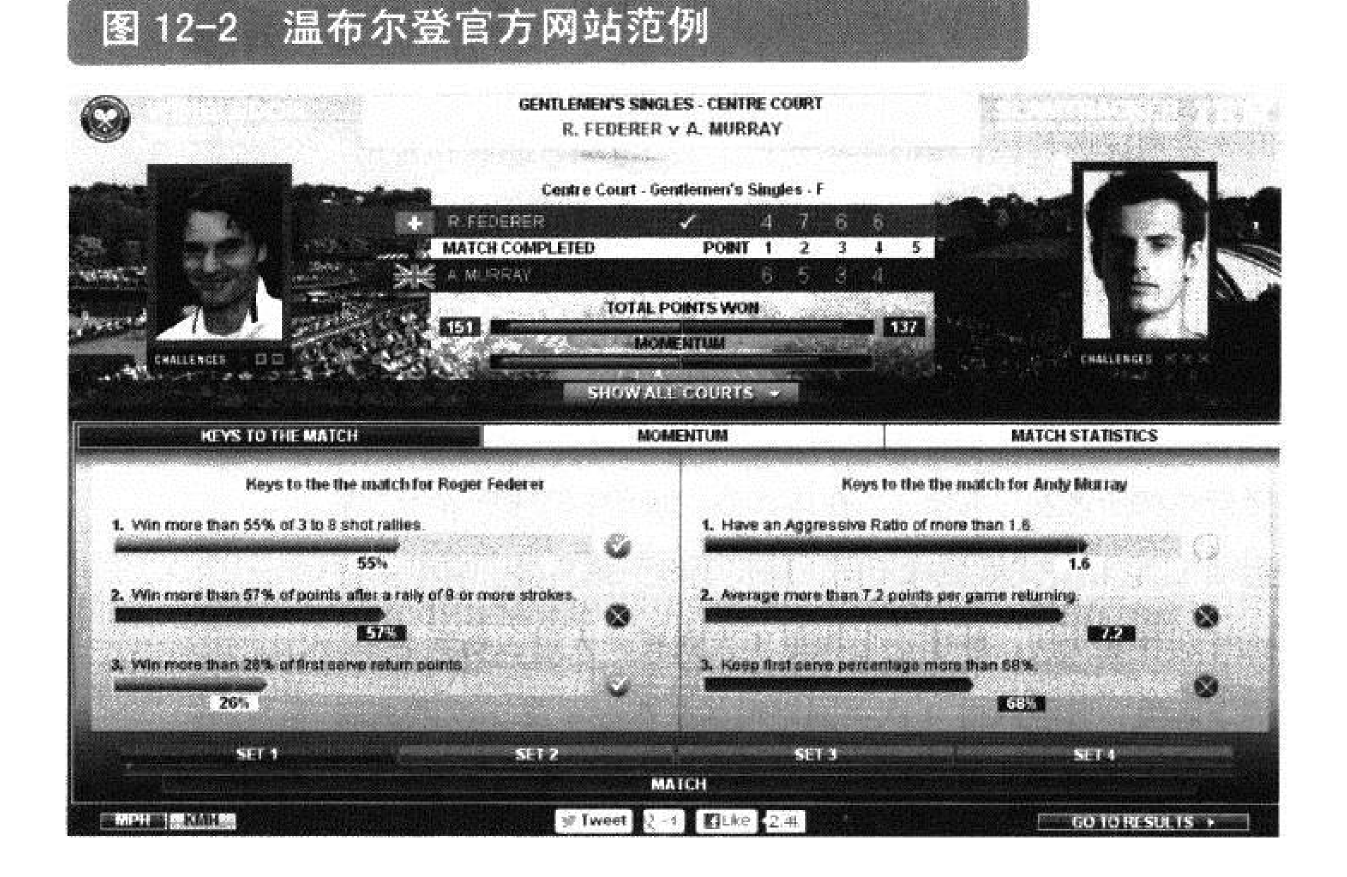
在举办期间网站浏览需求都非常高，共享云资源等于可以创造出经济规模，既能使用最先进的技术架构，又可减少各主办单位投资和管理的负担。

除了对主办单位方便外，当然网站最重要的还是要能吸引球迷。进入温布尔登官网，球迷可以浏览基本的赛程和计分信息，以及了解每位球员的详细数据，例如，在男子球员名单中找到台湾著名网球运动员卢彦勋，点入即可看到他历年的单双打战绩和累计奖金金额，还可观赏在线同步转播的精彩画面回放，实时了解他的表现和积分。

外行看热闹，内行看门道，但不管是凑热闹或懂门道的人，大会都想为他们创造更独一无二的体验。因此，球赛每一场、每一盘、每一局、每一分的数据都被收集进来，每个得分都会产生几种不同的记录，包括发球的速度、网前次数、发球犯规次数、是反手或正手拍回球等。官网会实时收集和显示这些数据，转播单位可以立刻用来评论赛情，教练可以分析选手的表现以实时调整对战策略，世界各地的球迷则可通过计算机或智能手机的应用程序追踪选手的表现。

在赛情实时数据外，IBM 事前还整理了过去 5 年大满贯赛的资料，把总共 3900 万条的数据汇入数据仓库内，再运用统计分析工具，预测每一个球员需要达到什么技术指标，包括接发球点胜率、每盘破发成功率、网前得分率等，才能提高赢球率。比赛进行间，球员在场上的表现会和这些指标交叉比对，例如，点入男单冠亚军费德勒对莫瑞之战，便可看到一发进球数、ACE 球数、上网得分率等数据（如图 12-2 所示），增加比赛的可看性，也可加深球迷对网球运动的了解。

图 12-2 温布尔登官方网站范例



减少营运风险

2012 年 4 月，带领纽约尼克队过关闯将、掀起“林来疯”热潮的林书豪因为膝伤开刀，整个季后赛泡汤。除了球迷失望外，最头痛的应该是尼克队和 NBA 联盟，媒体评估，少了林书豪，球队和联盟将减少数百万美元的收入。

一个球员受伤，为什么会造成这么大的损失？答案很简单，职业运动不仅是一门运动，其实更是一门生意，而且是很大的生意。例如，组



约尼克队年营收超过两亿美元，不仅 NBA 的明星球队，职棒大联盟（MLB）顶尖球队每年的营收也在数亿美元左右。但赚得多，花得也不少，球队不但要管理球场营运，聘请球员和员工，还得经营品牌和营销，样样都得砸大钱。万一有球员受伤，特别是像林书豪这样的明星主力球员受伤时，可能会影响球队在场上的表现，以及冲击票房而造成严重的损失。

正因为如此，英国的职业橄榄球队莱斯特老虎队（Leicester Tigers）想要尽可能降低运动伤害发生的概率。在体育界，英式橄榄球被称为“绅士玩的野蛮运动”，它和美式橄榄球不同，球员上场不戴头盔、肩甲和臀垫等护具，也特别容易受伤，球季中平均每 4 人就有一人进入伤兵名单。老虎队尽管战绩辉煌，曾经赢得 9 次英国橄榄球联赛冠军及 2 次欧洲联赛冠军，但也同样面临球员受伤的高风险。

每 4 人就有一人受伤，球队损失恐怕相当惨重，该怎么办？最好的办法就是好好保护球员，根本别让他们受伤。老虎队采用 IBM 的预测分析软件，通过深度的分析评估球员的出赛密集度、训练量和疲劳指数等，当指数产生明显的变化时，就可以提前预测运动员可能受伤的程度升高，球队即可因此调整训练策略，以降低球员受伤的可能性。

老虎队共有 45 个队员，运用这套系统，球队可以更了解对每个球员而言，哪个指数最可以预测出他们可能受伤的程度。球员进行训练和练习赛时，可在肩胛骨间贴上无线传感器测量压力和疲劳指数，若感测到球员的疲劳指数已接近他可承受的上限，表示再继续下去可

能造成运动伤害，教练便可设计更个性化的训练计划，缓解球员的肌肉疲劳。

老虎队也可分析哪些心理因素会干扰球员的表现，譬如客场比赛的压力比主场大，或者是否有其他社会或环境因素会增加球员的心理压力，而导致他们更容易受伤。了解这些因素后，球队便可设法对症下药。另外，老虎队也把这套分析机制用在19岁以下的青年队球队中，衡量未来生力军的压力和疲劳指数，作为挑选老虎队正式球员的参考，并且找出对球队最有利的球员阵容。

精准营销

成立于1948年的WAZ媒体集团（WAZ Media Group）是欧洲一家大型的媒体集团，旗下有13家媒体公司，约有1.5万名员工，出版近300种刊物，包括27份报纸、13本周刊、175种消费和技术性杂志以及99种免费报纸。

WAZ是德国颇负盛名的报业集团，但这几年印刷媒体的日子越来越难过了，网络上的免费新闻随手可得，消费者不愿意掏钱买报纸，广告主跟着流失，最主要的财源广告收入也大幅萎缩。WAZ认为必须改变营销方式才能引读者花钱订阅，所以，该公司建立一套客户关系管理和分析的系统，通过评估客户的行为和喜好，推出更具针对性的营销方案。

这个方案采用和温布尔登网球公开赛官网一样的分析软件，当系

统从不同来源采集 WAZ 现有和过去的订户记录，以及客户身份背景等信息后，可以在很短的时间内分析大量的数据，找出客户之间的共同点。

营销人员能够清楚地掌握在已推出的广告和促销方案中，哪些收到客户的响应、不同客户各通过哪些渠道响应、各目标客户族群最容易收到哪种促销的吸引、哪种订价方式能吸引到哪些客户。他们还能使用系统的演算模型分析几种客户的购买行为，包括现在的订户、过去的订户，及曾买过集团刊物并留有记录的客户，以预测这些客户群体未来的采购决定。

对营销人员来说，掌握这些信息非常重要，因为 WAZ 旗下有上百种刊物，照理来说，买商业管理月刊和买青少年潮报的读者应该很不一样，但以前营销人员只能猜测每种刊物读者群大概的特性。

现在则不同了，他们能按照常用的沟通渠道（如 DM 或电子邮件）、年龄、性别、婚姻状态、收入和住址，把客户分成不同的目标族群，评估每种刊物读者的特色，并且比较不同刊物读者之间有什么共同点，例如，某财经日报和高尔夫球杂志的订户群很类似，便可针对这两种刊物的读者进行交叉销售。

而且，他们还可分析退订杂志和报纸读者的数据，了解客户不再续订的原因，如果发现其中有些共同的原因，就能加强服务有类似情况的现有订户，或提供更个性化的优惠方案来吸引他们续订。通过精准的预测和分析，WAZ 可在客户退订前提早反应，已经降低了客户流失率，并把客户对促销方案的响应率提高到比业界高出三成之多。

延伸应用

随着互连消费者时代来临，媒体与娱乐企业必须师法一般 B2C（business-to-consumer，对消费者提供商品或服务的经营模式）企业，更直接与消费者互动，积极寻求并采纳消费者的想法才能成功。

通过大数据分析，企业可更精准地预测消费者动向，如 WAZ 媒体集团分析读者退订的原因以留住客户，也可发现新市场和开发新产品及服务，如温布尔登网球公开赛在官网提供深入的观赛体验，或如莱斯特老虎队，通过分析球员的身心状态，提早感测并预防风险。

大数据分析还能够帮助媒体与娱乐企业更有效地管理信息，在制定决策、策略和开发内容时做出更好的预测，并且设定备选方案的优先级。更明确地讲，媒体与娱乐公司可进一步运用客户分析：

- 汇总产业生态体系中多种内容供货商、合作伙伴及非合作伙伴平台中的消费者个性、习惯及喜好优先级的信息；
- 整合过去和现有的数据，以提高客户分析的使用效率；
- 根据实时客户分析，分析客户消费状况和广告活动之间的关联；
- 针对各种不同频道、平台、通路和设备，提供个性化营销和销售活动。

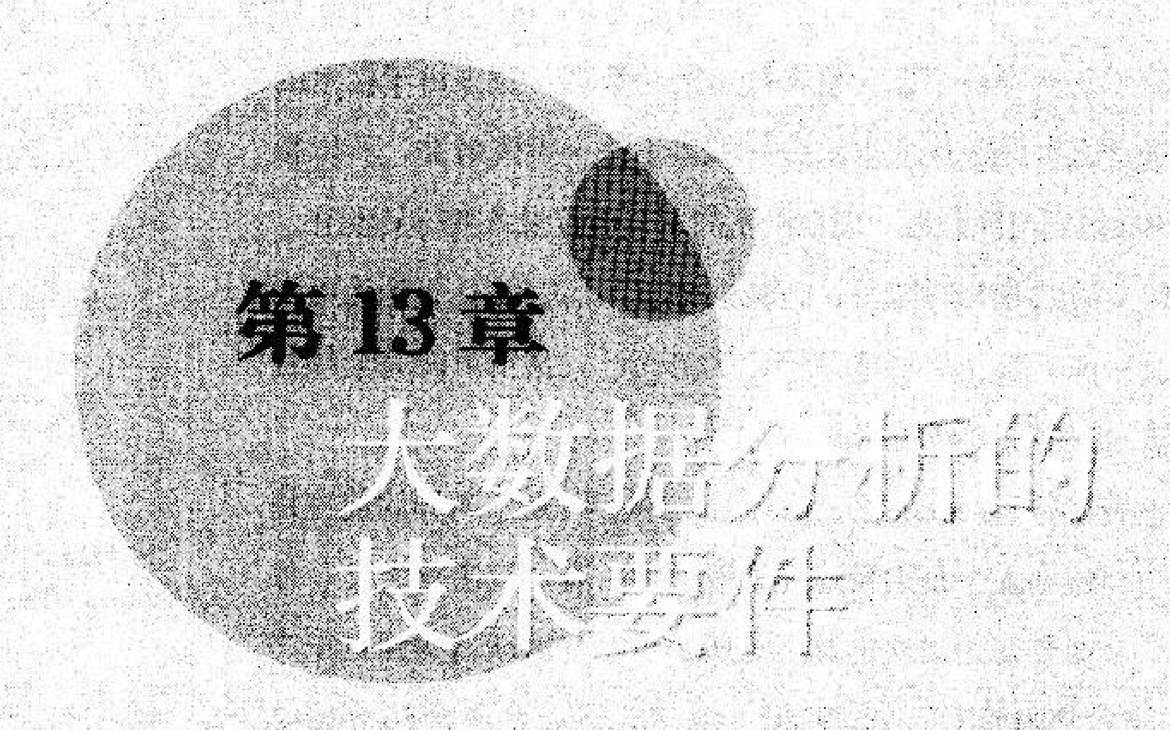


Part 3

技术与前瞻


pin51.com
拼吾爱
最新编程资源分享下载站





第13章 大数据分析的技术要件

前几章中，我们介绍了大数据分析的价值，研究人员可以在海地地震灾后实时锁定传染病疫区，阻止病情蔓延；销售专家可以重组商品排放在零售货架上的位置和售价，减少 17%的存货；开采油井的工程师可以调整钻油的位置和停机的时间，以提高超过 5%的产量。

这些仅是一小部分的例子，如同我们一再强调的，分析正是化数据为金的关键！

不过，组织中仍然存放了极大量未被运用的数据，过去由于技术的限制，想要分析这么多的数据非常昂贵，根本不符合投资效益，所以很少有企业或组织做得到或愿意做。现在则不同了，储存设备变得更便宜，分析的技术变得更成熟和相对普及，大数据分析的潜在效益更是无可限量——这个效益可能不仅止于创造差异化优势而已，甚至可能完全颠覆产业运作的本质，开创出全新的天地。

看看百年老店奇异（General Electric，GE）的例子吧。奇异的前身是爱迪生电灯公司，创办人正是名闻遐迩的爱迪生。有“交流电之父”作开国元老，这个企业集团的表现也算不负所望，总产值就占美国电工产业产值的 1/4。奇异集团旗下的奇异能源（GE Energy）在全球 50 多国设有几千台燃气涡轮机组，通过安装在机组上的传感器，每天不停收集机器震动和温度变化的数据，以确保机组运作正常。对他们来说，大数据早已存在多年，绝非最近几年才出现的新鲜事。

不过，最大的差别是，他们的储存和分析计算能力变强了。

以前系统只能存放 3 个月的旧数据，以及分析少数几个关系数据库

内所储存的数据，所谓的"分析"，多半只是在数据库中输入条件，查询机组作业的历史数据中是否有类似的情况，好像把数据库当成字典在里面查单字而已，但就连这么简单的作业都得花上十天半个月。

现在则不同了，他们建立新的平台，运用数据压缩和数据分析技术，工程师不但能够立刻查询10年前的数据，还能同时执行上百个演算程序，从历史数据和实时作业信息的交叉比对中解读出意义，以预测零件是否有故障的迹象，在问题发生前先行调配和修复。通过这样的方式，奇异能源可以不中断地供电，甚至不必再因为年度设备检修而暂停服务！

奇异能源公司的经验告诉我们，强大的平台有助于加强分析能力。那么，什么是平台呢？在信息的世界里，平台（platform）泛指涵盖硬件和软件的架构，其中软件也包含应用程序和操作系统，而软件（特别是应用程序）则会在平台上"跑"（执行）数据。传统上，应用程序和操作系统会针对特定平台架构所设计，因此，如果要跑某些应用程序，通常就得采用特定的平台才行。

壁垒分明的智能手机平台就是一个很好的例子。智能手机平台的两大山头是谷歌研发的开放式平台 Android，以及苹果的封闭式平台 iOS，各有不同的软硬件规格。硬件包括手机的处理器架构，软件则是大家常说的手机 APP，APP 是 Application（应用程序）的简称，像是风靡全球的游戏愤怒的小鸟（Angry Birds）或通信软件 Whatsapp 都是 APP。

APP 必须符合平台的规格才能执行，因此，程序开发商必须针对个



别平台设计不同的版本，这也是为什么你的 iPhone 和你朋友的 HTC 手机能下载的 APP 不太一样的原因。以 Angry Birds 来说，开发商的芬兰 Rovio Mobile 公司一开始只推出 iOS 的版本，所以只有 iPhone 用户玩得到，后来随着这款游戏受欢迎的程度增加，才又推出支持 Android 平台的版本。

因为平台会决定你的手机能够执行哪些程序，所以，选购手机时必须考虑要用哪种平台，同样的，企业要处理和分析数据，也需要选择适当的平台。尤其当你面对的是大数据时，数据的源头和种类更多元、数据的变化速度更快，更需要有一个平台处理不同型态的数据，以便于开发和应用数据。

另一方面，分析平台也提供一般用户接触和处理数据的接口，如果设计得简单方便，让就算不是计算机工程师的人，如职业篮球队的教练、保险公司的理赔专员、新生儿加护病房的护士等，都不需要靠专家协助就能跑分析的话，分析的效益才能真正发挥出来。

大数据分析平台六要素

大数据分析平台上跑的是大数据，由于大数据更大量、形态更复杂，变化也更快速，所以，和传统“不大”的平台相比，大数据分析平台所需应用的技术也不一样。

有什么不同？想要回答这个问题，我们得先回到第 2 章谈过的大数据的 4 个 V——大、杂、快、疑。



- Volume："大"量存放中的数据（data at rest）。
- Variety：种类繁"杂"的数据（data in many forms）。
- Velocity："快"速变动的数据（data in motion）。
- Veracity：真伪存"疑"的数据（data in doubt）。

我们先来看前 3 个 V，每个 V 分别代表分析大数据时会面临的数据源，各自能以不同的技术来对付，如图 13-1 所示。

图 13-1　3 种数据形态及对应的分析技术











要素 1：数据仓库技术（处理第 1 个 V：“大”）

第1类数据是结构化数据，来自于传统的数据源，包括企业的 ERP、CRM、SCM 和人力资源管理等应用系统，以及支持日常业务营运的核心系统，如负责管理和处理存款、贷款、汇款和支付等金融产品的银行核心系统，或覆盖承保、核保、理赔、保安等各个环节的寿险业核心系统等。

这些系统产出的结构化数据保留在关系数据库内，按照事先设定的格式（结构）所组织。但一个企业可能同时拥有好几个数据库，数据库如果各自独立，数据被拆散在不同的数据库里，就很难拼凑出营运的全貌。

以银行的信用卡电话营销活动为例，银行的 CRM 数据库中存放着销售人员团队与客户互动的记录，包括客户什么时候购买什么产品等，假使客户是因为某次促销活动而下单，CRM 系统也应记录下来，而这笔订单就是当时谈成生意的 A 业务人员的业绩。同样的，银行的人力资源管理系统也应记录这一笔订单，作为 A 年终考核的基准。

虽然 CRM 系统和人力资源系统都有和这笔订单相关的记录，但是，数据的内容和格式却有差异，CRM 系统数据库内有着和同一位客户接触的详细记录，包括 A 和客户通话多久和几次、A 什么时候发短信给客户、客户申请的是哪一种信用卡等，以及除了 A 以外，还有 B、C、D 等营销人员联系过同一位客户的消息。至于人力资源系统则记录 A 这个月多了



这笔业绩、可换算成多少奖金。

注意到了吗？A 在谈成这笔生意前的活动，以及 B、C、D 所付出的时间成本并没有列入人力资源系统，而 CRM 系统也不会记录这笔订单为 A 带出多少年终奖金。但银行必须要掌握每一笔数据，才能了解公司为这笔订单究竟投入多少营运时间和资源，也才有机会找到改善营运效率的办法。

此时，数据仓库便派上用场了。我们可以把 CRM 和人力资源系统的数据库想成两个实体的小仓库，里面都有货架，CRM 和人力资源数据则放置在货架上，各自依照编号排放。数据仓库则像是一个中央的大仓库，按照事先设定的规则，从小仓库的货架上提取不同编号的数据，这些数据如同工业生产的原料，会被加工处理、整理和转换成可以分析的格式后，再运用分析的软件“生产”出成品，也就是可用的信息。

由于第 3 章已经介绍过数据仓库和数据挖掘，所以这里就不再赘述。但值得一提的是，数据仓库的技术已相当成熟，市面上也有不少数据仓库的方案能帮助企业分析大量的结构化数据。

要素 2：Hadoop（处理第 2 个 V：“杂”）

第 2 类数据则是半结构化与非结构化的数据，来自非传统的数据源，包括网页内容、社交网站、电子邮件和单击流等，这些数据的格式与结构化数据不一样，数据格式多元且繁杂，却占据全世界所有数据的 85%。针对大数据这个“杂”的特性，Hadoop 是最适合的分析技术。



Hadoop 是由 Apache 软件基金会（Apache Software Foundation）所研发的开放源代码（open source）分布式计算技术，是以 Java 语言开发，专门针对执行大量且数据结构复杂的大数据分析所设计，其目的不是为了瞬间反应、提取和分析数据，而是通过分布式的数据处理模式，大量扫描数据文件以产生结果。

或许你正好奇"Hadoop"一字从何而来，有趣的是，Hadoop 并不代表任何英文字或缩写代码，只是一个无中生有创造出来的名称。当初原始开发者库廷（Doug Cutting）在为这个新技术命名时，他想选一个容易拼字和发音、便于沟通，且没有在其他地方使用过的代码，于是神来一笔地借用儿子黄色绒毛填充大象玩偶的名字，而黄色大象后来也变成 Hadoop 的官方吉祥物（见图 13-2）。

图 13-2　Hadoop 官方标识



Hadoop 不需使用商用服务器，在一般个人计算机（PC）上就能运转。用户可利用网络连接两台以上的 PC 组成服务器群，即所谓的"集群"（cluster），集群内的主机会分工合作处理数据。随着要处理的数据量越来越大，只要不断增加计算机的数量，而不需修改应用程序代码，就能立刻增加 Hadoop 的计算能力。

万一数据真的很多，就得买很多台 PC，不是得花很多钱吗？这么说虽然没错，但看看一些数据，你就会发现 Hadoop 还是很"便宜"：买一

台效能为 PC 两倍的主机，你得付出远高于两台 PC 的费用；买大型主机得花上千万元，但同样的钱可买到数百台 PC，通过 Hadoop 整合便能提供超过一台大型主机的效能。

换言之，Hadoop 可以用更低的成本，得到更高的计算效能，增加数据分析的能力，也难怪有些人称 Hadoop 为大数据的救星，这种说法虽然夸大，但的确有几分真实，因为通过 Hadoop，就算荷包不够深的个人或组织，也能分析大量的结构和非结构数据！

如果是资本雄厚的大公司，就更有本钱大大扩充 Hadoop 的计算量了。美国 Yahoo!在 2008 年就宣布采用两千台服务器，把 Hadoop 技术应用于旗下的搜索工作上，例如比对同义字、热门关键词分析等。另外，沃尔玛、eBay、VISA、中华电信和台积电亦已借助 Hadoop 进行数据分析，包括用户行为分析、防堵诈骗及提高生产线合格率等，并产生显著成效。

除了“C/P 值”（性能价格比）很诱人之外，Hadoop 另一个优点是其相对稳定和高效率的数据处理模式。为了说明这一点，我们得看看 Hadoop 计划下的两个主要项目：分布式文件系统（Hadoop Distributed File System，HDFS）和分布式处理程序框架（MapReduce）。

● HDFS：数据切割、制作副本、分散储存

HDFS 会把一个档案切割成好几个小区块、制作副本，然后在 Hadoop 服务器群集中跨多台计算机储存副本。

首先，要处理的数据会被拆散成一个个档案区块（block），每个区块

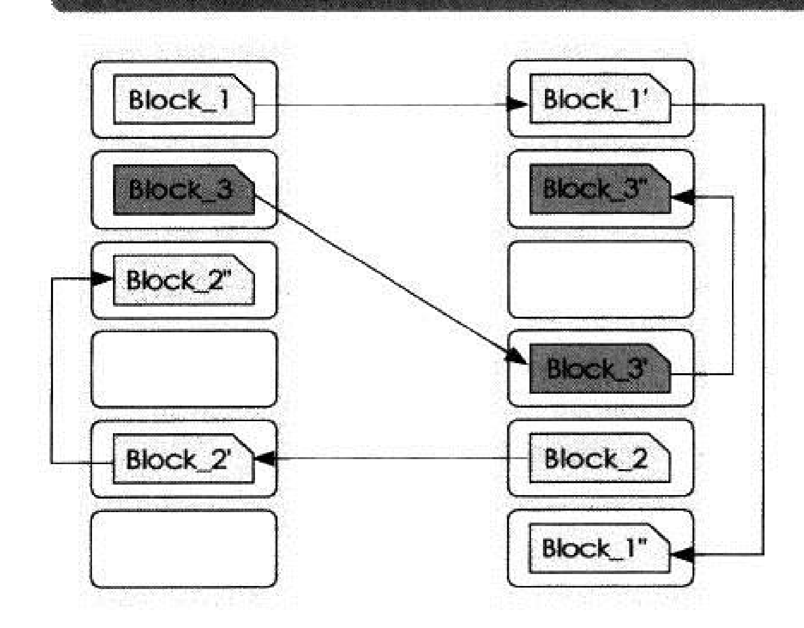
是 64 MB。例如，有个档案是“克林顿总统 2012 年演讲稿”，大小为 180 MB，便会被分成 3 个区块（见图 13-3）。

Block_1：64 MB

Block_2：64 MB

Block_3：52 MB

图 13-3　HDFS 复制档案区块的方法



接着，系统会自动制作副本，预设是 3 个副本，如第一个区块的副本是 Block_1、Block_1'和 Block_1"，依次类推，每个副本会被存在 Hadoop 集群里不同的计算机上。在这个例子中，集群里共有 12 台服务器，这 3 个区块共 9 个副本存放于 9 台不同的计算机中。因为一台计算机只能储存一个区块，所以在放入 9 个副本后，这 9 台计算机就不能再存放其他区



块的副本了。

● MapReduce：拆解任务、分散处理、汇整结果

MapReduce是由Map和Reduce组成，前者分散计算数据，后者则负责汇整Map计算完的结果并输出。

MapReduce的运作方式很像是超市盘点存货：店经理先将盘点任务分配给不同类别货品的部门，每个部门各自负责所属的货架，有人负责清洁用品、有人负责冷冻食品、有人负责生鲜食品……大家完成统计后，各自将存货量回报给店经理，最后由店经理统一汇整出整家店的存货。这样就不需要由店经理一个人单打独斗逐一清点货架的商品量，而是通过分散处理的方式来加快盘点作业。

盘点任务就像是Map程序，每个部门都执行同样的盘点程序，并只负责处理局部数据，而Reduce就是汇总存货量的工作。

还是回到“克林顿总统2012年演讲稿”档案的例子。当用户要求MapReduce执行任务，如计算讲稿中提到“社会保险”的次数，系统会在每个区块的3个副本中各挑一个，例如挑出Block_1、Block_2'和Block_3"，要求它们统计各区块内“社会保险”出现的次数。

还记得吗？这3个副本组合起来就是原先180 MB的“克林顿总统2012年演讲稿”档案，被分成3份储存和计算。换句话说，本来一台计算机得处理180 MB的大档案，现在则交给3台计算机，分别处理64 MB、64 MB和52 MB的小档案，等于每个副本只需做本来工作的1/3就好，



所以速度当然可以快很多。

更厉害的是，当某个副本损坏时，MapReduce 还会自动侦测，改派另一副本执行任务，例如，Block_1 坏掉时，系统会从 Block_1'和 Block_1"间选出“继任人选”。因为 Hadoop 一般是在 PC 上运转，PC 的故障率比商用服务器高出许多，所以，这种“容错”的功能非常重要，当集群中有 PC 损坏时，才能继续执行任务。

简单来说，Hadoop 借由把数据切割、分散存放和处理的方式，让集群内每台计算机只需要处理小部分的任务，大大提高数据分析的效率，再加上可以同时处理结构化和非结构化的数据格式、相对便宜的建立成本及容错的特点，成为大数据分析很重要的技术。

要素 3：流计算（处理第 3 个 V：“快”）

大数据的第 3 个 V 是 Velocity，实时变动的流动性数据（data in motion），这类数据产生后，基于某些业务需求，只有极短的时间可以处理。例如，在 9.11 事件发生后，美国政府加强在机场港口等地的入关安检措施，运用脸部识别技术，必须在旅客通关的短短 10 秒内找出通辑的恐怖份子，其分析就分秒必争，非常讲求时效性。

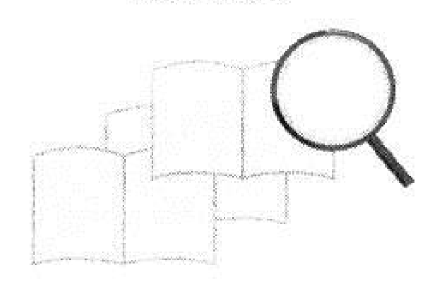
这种异动频繁、流量极大，又需实时响应的数据，已经超过传统的数据库管理模式的处理能力，而得交给流计算处理。流计算是源自于 IBM 与美国国防部合作研发的反恐系统，能针对大数据进行实时性、高复杂度的分析，最快可在微秒内做出反应与决策。

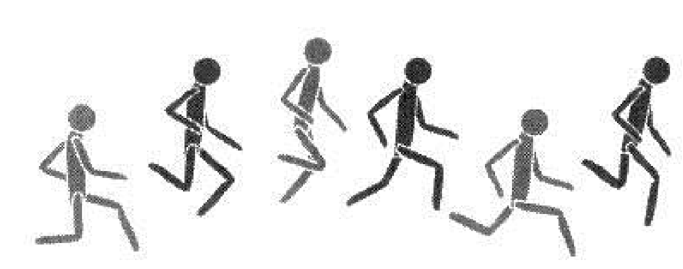


图 13-4 流计算对比传统数据库管理模式

传统计算模式



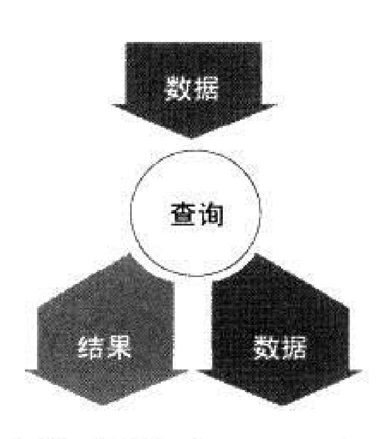




在已存放的历史数据中找答案

即时分析持续流动的数据





传统的数据分析模式中，数据得先收集到数据库中再搜索或查询。举例来说，企业的人事数据库定期进行批次性的数据统整，里面存放着与所有员工相关的信息。若要统计或分析数据时，必须发动查询的任务，如“2012 年 12 月 31 日为止，工龄满 5 年的员工”，数据库引擎便会在既有的数据中搜索，下钻出符合查询条件的员工数据。

数据库的引擎平常不会运转，只有在接到查询指令后才会启动，启动后也是“一个指令一个动作”，收到什么任务就交出什么样的结果。

流计算就不同了，流计算有个不断运转的引擎，进入引擎的不是查询指令，而是源源不绝的数据流，可在数据储存前先完成分析。

我们可把汇入流计算的数据流想成一列骨牌，推倒一张后，后面的牌就会一张接着一张产生连锁反应而倒下。好比海关安检，数据产生（摄像头拍下旅客脸部照片）后流入引擎，触发计算引擎将影像转成可判读消息如旅客脸部特征点（眼口鼻的位置和距离、黑斑、痣等），把这些特征点和各安检数据库中的数据比对之后，再拉出特征点相符之列管恐怖份子的照片。

这只是一个旅客而已，想想看，美国的大型机场每小时有几千人入境，全美约有 1000 个国际机场以及几百个港口，同时读取、转换、比对和判读旅客数据是多么复杂又庞大的工程！流计算却能动态收集多个数据流，使用先进的算法将结果实时传达给决策人员，因此能被应用于更复杂、更机动且须立即决策的数据。

要素 4：数据治理（处理第 4 个 V："疑"）

上面 3 个要件处理的是大数据的前 3 个 V，现在，轮到第 4 个 V 了，也就是不确定性（Veracity）。

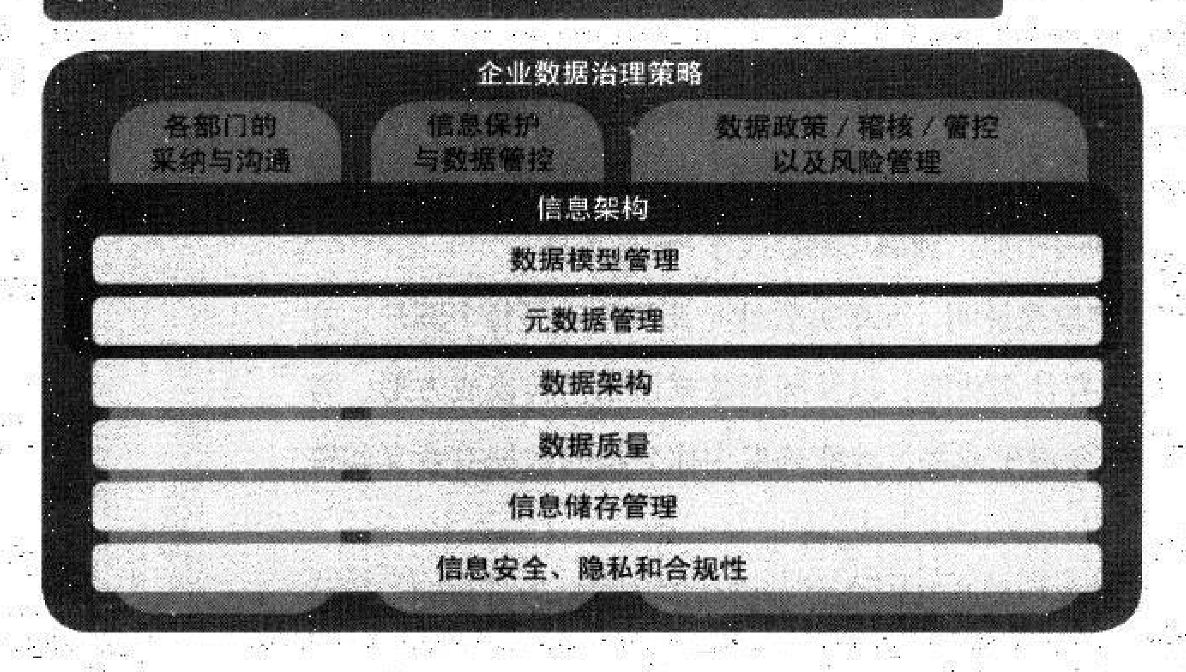
大数据的源头远比以前多，除了传统上由企业信息系统所产生的数据外，还有越来越多来自社交网站和传感器等非传统源头的数据，使得数据真伪难辨、破碎不全的比例越来越高，但数据一定要可靠和完整，分析才有意义，因此，第 4 个要件"数据治理"虽然不负责数据分析，却可能是影响大数据分析成败的关键。

什么是数据治理？"治理"（governance）和"政府"（government）的英文都有 govern（管理、监督），两者在意思上也很相近。把治理的概

念运用在数据上，指的是组织要像政府管理政务般，通过一系列的措施与程序、详细设定管控的机制，妥善地管理并监督数据。

数据治理可以分成几个层面来谈，如图 13-5 所示。

图 13-5 数据治理架构



1. **数据模型管理：**不同的系统有不同的模型，当数据从上游进入下游系统时，必须转换成一致的模型才能进行处理和分析。举例来说，同一笔“A 连锁快餐店当日营收额”的消息，门市的销售点系统可能依照各产品的营收贡献度来分类呈现（如薯条销售额占营收 10%、汉堡占 50%、可乐占 10%等），而物流系统则按照各门市营业额来保存（如东区门市营业额 150 万、西区 200 万、南区 120 万等）。这两个系统描述的是同一件事情，却以两种形式呈现，这两种形式就是两个不同的数据模型。



看过烤蛋糕的烤模吗？这两种模型就好像一个圆形、一个方形的烤模，“A 连锁快餐店当日营收额”则好比一个面团，放入烤模加热可变成不同形状的蛋糕。而数据模型管理的任务就是要知道组织内总共有 100 个圆形、20 个方形、40 个星形的烤模。

2. **元数据（meta data）管理**：元数据是描述数据的数据，元数据负责定义数据和规范数据间转换的定义。以快餐店的营收为例，其元数据可能包括什么叫做营收（如各门市销货收入），以及营收如何转换为利润（如：利润=营收-成本+利息）等。

同样以蛋糕烤模来说明，元数据管理是要清楚掌握每个圆形、方形和星形烤模各有什么功能，以及圆形数据该如何转换成方形、方形数据又要如何变成星形等，才能确保用作分析的数据有意义和符合需求。

3. **数据架构**：数据架构是数据从源头到最下游分析的过程间，经过多次整理的过程，譬如，快餐店各门市营收→各区域营收→公司总营收→公司利润，各阶段都要有明确的组织和架构，以保障分析数据正确可靠。

4. **数据质量**：数据质量指的是数据正确、完整且可被有效率地使用，市面上已有许多可衡量、改善和验证数据质量的方案。

5. **数据储存管理**：好的储存管理有助于加快数据处理的速度，一般而言，数据需按照使用频率和价值区分为不同层级，刚“出炉”热腾腾的数据及经常使用的数据应存放在反应速度较快的储存设备上，以方便存取和提高效率，而比较少使用、比较“冷”的数据，则可



储存在比较低级或反应较慢的设备里，以节省储存设备的开支。

6. 数据安全、隐私和规章：最后，还要保障数据的安全和隐私。不是每一笔数据都应该拿来分析，也不是组织内部每个人都有权接触所有的数据，这些原则都需要有好的规范和机制辅助才能落实。

要素5：文本分析

大数据充满各种社交网帖子、电子邮件和新闻信息等文本，对传统计算模式来说，这些非结构化或半结构化数据就好像是从未学过的外国语言一样，根本看不懂。更不用说该如何解读出意义了。通过文本分析，这些原本“无意义”的信息则可从原始文字数据中被抽取出来，如同经过翻译一样，因而产生意义。

简单来说，文本分析是从文字数据提取信息的过程，有几种用途：信息摘要（从大笔信息或单个文件中找出关键消息）、消息分类（分析文本中提及哪些主题）和更复杂的情绪分析（如从网络的评论文章中分析消费者对产品的评价）。

下面这个例子是简单的消息分类过程（见图13-6）。

左边方格里有两段文字，属于非结构化数据，用户可通过文本分析，提取出文中提及的人名、职称和组织，并排放到正确的字段中，转成右边表格中的结构化数据。

完成这样的任务，系统需要有好几种词汇表（glossary），包括人名、组织名和职称名的词汇表。每一个词汇表就像黄页一样，罗列着不同的人名、



组织名和职称，当文本出现某个词汇时，如“Bill Gates”，系统就会去词汇表中检索、确认“Bill Gates”是人名后，就把它放入右边的“名字”字段中。

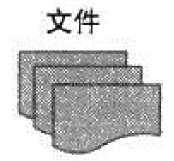
图 13-6　文本分析简例

For years, **Microsoft Corporation** *CEO* **Bill Gates** was against open sourse. But today he appears to have changed his mind. “We can be open source. We love the concept of shared source,” said **Bill Veghte**, a **Microsoft** *VP*. “That’s a super-important shift for us in terms of code access.”

Richard Stallman, *founder* of **the Free Software Foundation**, countered saying...

名字	职称	组织
Bill Gates	*CEO*	**Microsoft**
Bill Veghte	*VP*	**Microsoft**
Richard Stallman	*Founder*	**Free Soft.**

提取的信息



当然，不是每个文本都这么直接了当，组织或企业分析文本也不仅止于要分门别类而已。随着移动设备、社交媒体等新兴技术的普及，消费者越来越习惯在网络上分享消息、倾心吐意，企业也试图解析这些每天大量产生的非结构化数据，期望从中挖掘更多有关消费者对商品和品牌的观感，以调整产品开发和市场营销策略。

针对这类需求，文本分析还可升级到更复杂的语意分析（semantics analysis），从文字中解读发帖者的情绪或好恶。例如，我们在第 3 章介绍过，解析 tweet 的情感倾向可预测股价的变动，而且预测的精准度丝毫不逊股市高手。以 Facebook 挂牌当天的股价为准，开盘后 tweet 情感转向

负面时，半小时内 Facebook 股价开始下跌，而当 tweet 情感转正时，Facebook 股价立刻在 8 分钟内开始反弹。

下面的例子就更有趣了，分析样本同样是社交网站的帖子，不过，目的是为了预测电影票房。

钢铁侠、绿巨人、雷神索尔、美国队长、鹰眼、黑寡妇，这一长串奇怪的名字，你是否如数家珍？如果是的话，你应该看过《复仇者联盟》（The Avengers）这部 2012 年上映的超级英雄大片。但这没什么好讶异的，因为除了你以外，全球大约还有一亿人在电影首轮上映时挤进电影院，观赏这群超级英雄放下歧见、同心协力击退入侵地球的外星人。口碑和票房如此超群，也把这部电影推向影史上票房第 3 高的宝座，仅次于冠军《阿凡达》和亚军《泰坦尼克号》。

不过，在《复仇者联盟》上映前，有人已经预测到这部片的票房绝对不俗！

2012 年美式足球“超级杯”（Super Bowl）比赛期间，IBM 大量扫描 Twitter 和 Facebook 等社交网站和博客的帖子，从中搜集和分析网友对几部片商强打电影的评论，除了《复仇者联盟》之外，还包括《异星战场：强卡特战记》（John Carter）、《大独裁者落难记》（The Dictator）、《3D 恶灵战警：复仇时刻》（Ghost Rider: Spirit of Vengeance）和《海豹神兵：英勇行动》（Act of Valor）等片。

这不是 IBM 第一次把语意分析技术用在测知社交网站的“民意”动向上。在超级杯登场前，IBM 已和美国南加州大学合作发展出“奥斯卡民意量表”（Oscar Senti-Meter），根据 tweet 数量和 tweet 所表达的情绪，



为2012年奥斯卡奖候选的影片和影星排名，让全球的影迷都看得到哪部电影或哪位影星在网络世界中的人气和正面评价最高，例如，男演员中乔治克鲁尼最具人气，而麦特戴蒙则享有最多正面评价。虽然乔治克鲁尼最后没得奖，麦特戴蒙则连提名都没有，但分析结果的本身还是非常有趣，而且对电影制片公司来说，下次选演员阵容时，这些结果就有很大的参考价值了。

超级杯时，奥斯卡民意量表便担负起重任。其实IBM从2012年1月1日起，已经开始监测社交网站中对这几部电影的评论，不过最密集分析的时段是在超级杯举办的2月5日当天，IBM以每分钟为单位，记录网友提到每部电影的次数，以及对电影的好恶。

为什么选在超级杯？因为超级杯每年都会吸引一亿美国人收看，可说是全美年度最引人注目的活动，因此超级杯期间的广告效益特别高，好莱坞也都会抢在此时播出预告片，广告在网络上激起的讨论和关注特别多，正是分析“原料”最多的时候。

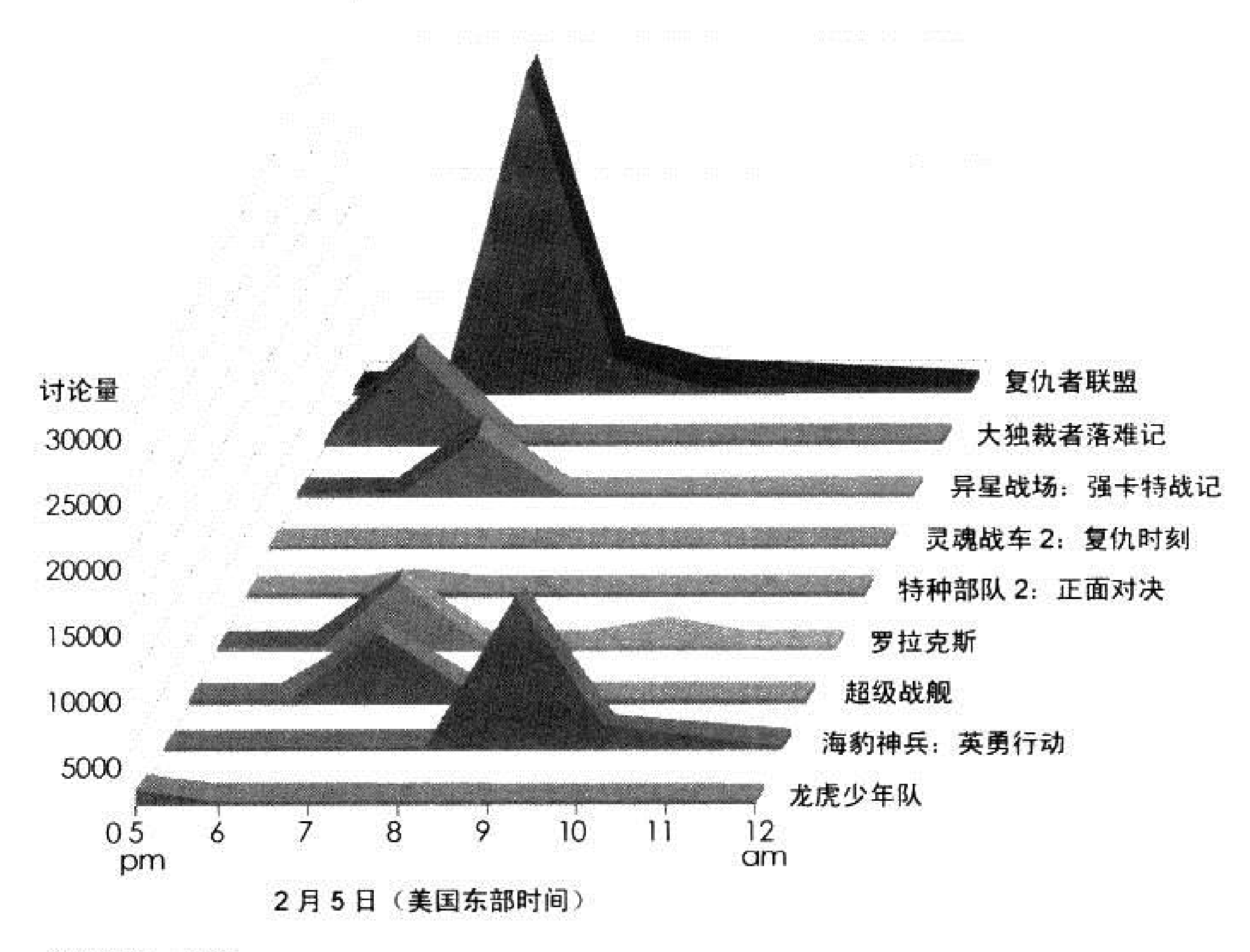
超级杯比赛共有4节，每节15分钟，在中间穿插的广告时段中，包括《复仇者联盟》在内的几部电影都播出长30秒的预告片。为了满足你的好奇心，顺便提一下，这30秒就要150万美元！

事实证明，超级杯的天价广告费物有所值。预告片播出前，网友对几部电影的讨论都近于零，播出后全都瞬间爆量：在晚上6:30～8:30比赛进行的两个小时内，IBM共收集到11亿篇tweet、570万篇博客和论坛帖子，其中有350万条消息和电影有关，谈论《复仇者联盟》的帖子数接近10万，数量足足比该时段上广告的其他电影高出好几倍，如图13-7所示。



图 13-7　以语义分析统计网友对电影的评论数

超级杯比赛期间，《复仇者联盟》预告片创造出远高于同期其他电影的话题



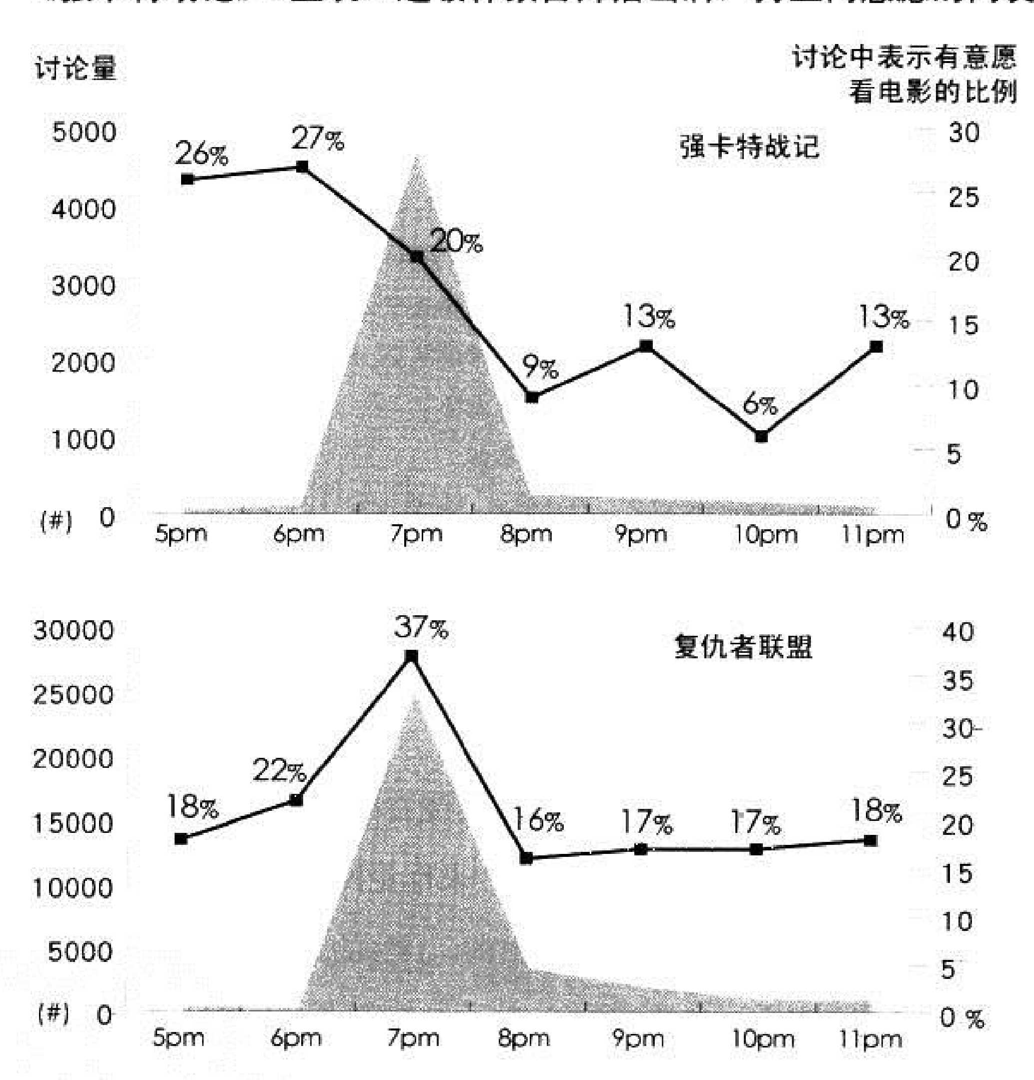
数据来源：IBM

预告片有人讨论当然好，但最重要的还是要让大家愿意掏钱看电影。于是，IBM 还运用奥斯卡民意量表分析网友情绪的正反向，来推论网友是否有意愿花钱看电影。分析结果显示，在超级杯广告播出前，18%的人对《复仇者联盟》有正向情绪，播出后达到35%的高峰，之后虽然下滑，但仍维持在接近18%的比例，表示将近两成的人有兴趣看这部电影。

另一部同样耗费巨资打造的动作大片《强卡特战记》就没这么幸运了。在超级杯预告片播出前，正向情绪的网友有26%，播出后居然不增反减，掉到20%，超级杯结束后更暴跌到只有6%～13%之间，显然当天的预告片做得不好，让不少本来有意愿看电影的人都转为负面（见图13-8）。

图13-8 从网络评论分析消费者看电影的意愿

《强卡特战记》（上表）超级杯预告片播出后，持正向意愿的网友比例大幅下滑

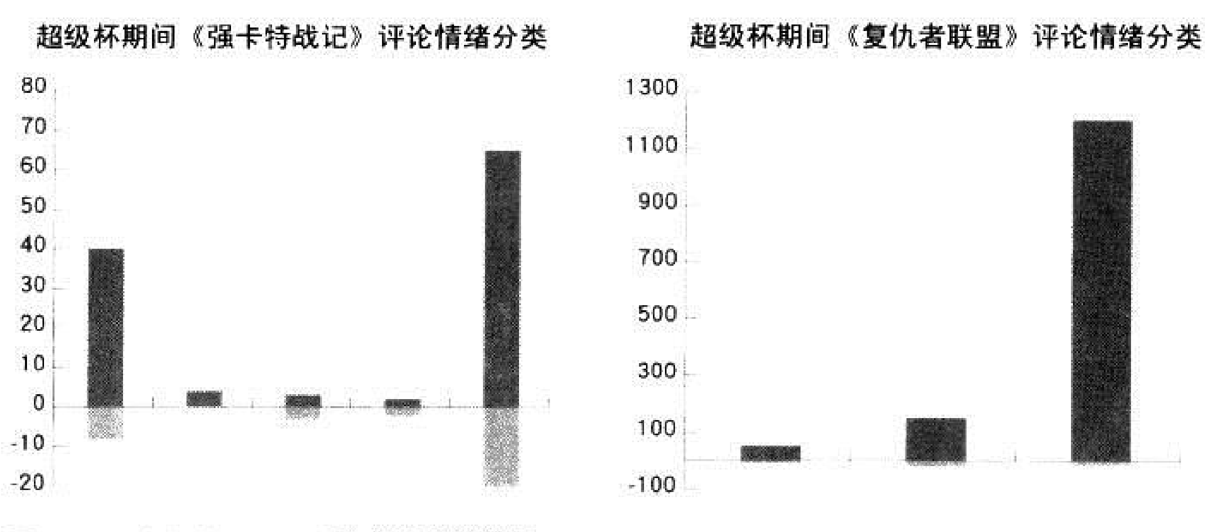


数据来源：IBM

除了统计评论数量和判断正负面意向之外，语意分析还能将评论内容按照议题分类，做出更细致的分析。IBM从提取出来的社交网络讨论中，进一步分析网友对《复仇者联盟》和《强卡特战记》在预告片、演员阵容、配乐、剧情和特效等5个项目上的观感，发现有关《复仇者联盟》的讨论绝大多数都集中在预告片上，而且几乎一面倒都是正面评论，而《强卡特战记》的演员阵容和预告片则得到较为两极化的评语（见图13-9）。

图13-9　按明细项目分析评论正反向观感

网友对《复仇者联盟》预告片几乎一面倒叫好



数据来源：IBM

电影上映后，《强卡特战记》在全美票房惨淡，几乎只有《复仇者联盟》的十分之一。对影迷来说，选错电影看，顶多就是花点冤枉钱、进电影院吹冷气睡上一觉，但如果你是电影公司的股东，票房惨遭滑铁卢恐怕会让你辗转难眠。



如果运用像奥斯卡民意量表这一类的语意分析工具，电影公司就能够在电影上映前先在网络上“试水温”，及早调整预告片的内容、配乐，或甚至重新拍摄电影的部分片段，尽可能提高票房，而创造出很高的投资效益。当然，除了娱乐产业外，社交网站的语意分析还可帮助很多公司开发消费族群、了解消费者对品牌和产品的好恶，而发展出无限商机。

要素 6：可视化和搜索接口

有了上述 5 个要素后，组织已经可以进行大数据分析了，不过，为了让一般使用者也能轻松了解、使用和查询数据，还需要提供简单易上手的使用接口，最基本的是具备数据搜索功能的查询接口，更高级一点则还可以图表等可视化方式呈现分析结果。

大多企业和组织内部已经有应用系统及结构纷杂的数据，当他们想要迈入大数据分析的领域时，第一步往往都是从查询界面开始。基本上，一般使用者虽然不懂信息工程，但多半很熟悉 Google、Yahoo!等搜索引擎；现在市面上查询的接口系统，使用起来和搜索引擎差不多，输入关键词后，即可同时搜索结构化、非结构化和半结构化数据，所以能够快速找到所需要的信息，并执行简单的分析。

珠宝品牌蒂芬尼（Tiffany & Co.）靠懂得消费者的心，叱咤精品界超过 175 年，他们新增的秘密武器，就是一个能让店员快速查询的搜索接口。每当顾客踏入店里，店员可通过一个简单的接口系统，查询顾客可能感兴趣的产品。店员只要输入顾客的特征，如“年纪看似 30～35 岁、

棕色及肩短发、着及膝无袖洋装与7.6厘米细跟包头高跟鞋、手拎Gucci包的女性"，系统会在庞大的顾客历史数据中做出交叉比对，跑出符合这些特征者可能感兴趣的商品，店员便可投其所好，立即为她推荐这些商品。顾客觉得店员了解自己的品味、懂得自己的需求，自然而然比较乐意掏钱血拼了。

有了这种容易使用的接口，组织内执行日常业务的人才愿意和能够接触所需信息，大数据分析也才真的有意义（见图13-10）。

图13-10　大数据分析平台的6大要素

还记得我们在第2章提过，大数据就像巨大的矿脉一样吗？大数据持续增加，如果依靠传统计算和分析模式，我们所能处理的比例越来越低，根本无法判断矿体里埋藏的究竟是稀世珍宝或弃土废矿。



但是，用对了工具，就能用符合经济效益的方式大规模开采。大数据分析平台就是这个工具，其 6 个要素解决了大数据应用和分析过程中不同的挑战和需求。组织可依据自己的信息策略、设定数据分析的需求，好像分层探勘矿藏一样，从最省力、明确的开采点切入，得到短期效益后，再逐步发展其他几项能力，更深入、更全面地挖出对加速组织增长、改善生活质量，乃至于促进社会发展有所帮助的宝贵发现，原矿也才能真正化为金！





第14章 结语与展望



当我们在撰写这本书的后期，一则国际新闻引起了我的注意。报导上说一位研究欺诈问题的数据科学家，发现欺诈的传播力和 DNA 的排序问题非常类似，在融合了两个完全不相干的世界之后，他和他的团队找到了一种能大幅降低欺诈损失的解决方案。

这代表除了消费者导向（consumer-oriented）的零售业、社交媒体，或者是之前我们所叙述的能源产业、健康护理和交通运输之外，未来这类跨领域的大数据分析实例会越来越多，而且几乎所有的产业都可以应用。

例如，我们知道某家世界级的轮胎大厂，已经在进行一项实验，可以针对车主的驾驶习惯、车型、环境等因素，探测哪一只轮胎何时该检修，同时告知车主离他最近的检修厂在哪里。

以前企业的思维是，希望驾驶员在使用轮胎的时候很安全，但是耗损率可以高一点，这样你才会在一定的时间内去更换轮胎。然而，这些坏掉的、报废的轮胎，焚烧了以后形成空气污染，掩埋了以后又是不可分解的永久性垃圾，对环境造成很大的影响。

但不久的将来，轮胎的使用可能会从按“数量”计费，变成按“里程”计费，利用车子上的传感器，可以预知这颗轮胎何时需要补刻胎纹，何时需要更换。如此一来，制造商的思维改变了，制作轮胎时会想要让这颗轮胎的使用寿命越久越好，消费者的思维也会跟着改变，那么我们对天然资源的消耗和污染的产生都会降低。

在中国，已经有很多城市开始利用跨领域的大数据分析，将我们的城市、产业、生活变得更具智慧。2013 年新北市就和全球其他 33 个城市一起，利用 IBM 提供的专业团队及技术协助改善市政，初期会以警政服



务为主，打造一个提升生活环境质量与城市安全的“科技防卫城”。

其中包括交通服务电子化、整合勤务指挥管制系统、广建数字影像监视系统、提高刑案侦查及鉴识专业能力等建构基础，以科技为主轴，通过有效的资源整合，提高治安服务效能。例如整合了原本各自独立的报案、警车卫星定位、地理信息、赃车车牌识别及监视器等 8 大系统的“情报整合中心”，就可以通过卫星、监视器实时有效地掌握案发现场状况，让警察赶到犯罪地点后，可以在最短时间内追踪歹徒逃逸方向。

另外，也有一些学术单位利用相关指针与数据库交叉分析，预测国内的社会消费品零售总额、城镇居民人均可支配所得、消费性支出与非消费性支出等发展，提供给零售企业用在销售据点选择、商品货色安排及预估销售量上。

这些都代表了各个产业正在借由大数据分析，超越并包含现有第一级产业（农、牧、矿）、第二级产业（制造业）、第三级产业（服务业）的限制，迈向以知识、智能为主的第四级产业（生活知识体验）和第五级产业（生命转化）。

知识＋智能：中小企业新出路

2013 年将是大数据的实践元年，尤以电信与金融服务企业最明显，伴随着企业应用案例的出现，也将会有越来越多的系统整合商（SI）与独立软件开发商（ISV）投入这个领域。

对于以中小企业为主体的产业结构来说，或许有人会认为，大数据



分析只适用于资金雄厚或数据庞杂的大企业。事实上，以目前的发展来看，大企业进入大数据分析的困难度的确较低，又或者是它需要的动力（motivation）较高，我们以台积电为例，由于它的制程数据是可采集的，所以进入的困难度较低，又因为是经营全球市场，加上只要汇率波动、运输状况等外部因素一变动，成本与获利会差很多，所以它需要大数据分析的动力自然也较高。

但是中小企业在这波数据浪潮当中，一样可以利用大数据分析，找到改变的新契机。例如，美国的奶农利用大数据分析搭配机器手臂来帮忙挤牛奶，传感器会记录每一头乳牛分泌乳汁的统计数据，经过智能手机上的 APP 上传到分析系统，就可以直接分析这些牛乳的生菌数，以得知乳牛是否健康，有没有感染乳腺炎，进而找出优化的挤奶策略，而机器手臂则会自动找出乳牛乳头，装上挤奶装置，这套系统目前已让一头牛可以多生产出 518 千克的牛奶。

另外，以色列的葡萄酒庄利用传感器，收集旗下签约葡萄园的数据，包括每一年的阳光、雨水、湿度、土壤里面的成分，进行大数据分析，来预测葡萄的糖分、酸味等产出情况，一方面提供给农夫作为种植时的参考，另一方面提供给酿酒师作为调味时的依据。

此外，它也把这些数据提供给品酒社群，让消费者共同参与葡萄酒的产制过程，甚至可以为客户提早预定这批葡萄要制成什么风味的酒款，这种做法让它的客户保留率是 100%，光是预购量就让这家企业不用再花钱做促销或广告宣传。

从这些例子中可以看到大数据分析应用面的广泛，而就中小企业居多的



台湾地区来说，如果国外的奶农和葡萄酒商能做，那么当地的茶叶、酱油、稻米等可不可以也用这个方法，进一步去提高产品的附加价值和客户稳定度?

如果一家小公司的数据完整性不足、经费不够，那么在产业转型的同时就应该思考如何发展一种新的商业模式，把该产业的数据和经费集中起来，形成一个可以共享的数据池（Data Pool），让大家共享这样的资源，来降低大数据分析的进入障碍，像是纺织工会或外贸协会就很适合提供这样的平台。

至于以代工为主的台湾地区制造业，也在“制造业服务化”的升级过程中面临无数挑战，运用大数据分析也能提供另一个不一样的发展策略，球承制造商瑞典 SKF 集团的例子就非常值得参考。

这家老牌的球承制造商，原本因为东南亚等地的同行低价抢单，加上北欧地区人工薪资高于亚洲甚多，使得营运成本居高不下，削弱了在市场上的价格竞争力，以至于订单大量流失。为了扭转劣势，SKF 在制造设备上装载了大量传感器，搜集有关震动、噪音等数据，希望借由提高制程及产品质量，来填补价格竞争力上的不足，没想到竟意外发现，在设备即将面临故障前，总是会出现一些共同症兆，SKF 也因此研发出一套可以预测系统可靠度的制程模型。

这项原本只是在内部使用的机制激发了 SKF 一个大胆的创新想法，SKF 认为球承制造过程中会面临的问题基本上大同小异，竞争对手也一样会遇到，何不把这套解决方案当作新产品来提供客户服务，协助同行解决生产设备的可靠度问题，增辟新的收入来源？果然这项新服务获得东南亚等地的竞争对手的欢迎，SKF 最后索性结束原本备极艰辛的产品



制造业务，彻底转型成为高价值服务的供货商，向原来的竞争同行提供制造优化建议及预警服务。

SKF 不仅利用大数据分析突破了原本制造业低价抢单的宿命，而且也以原本累积的知识和智能成功找到了新的出路，这也是大数据分析希望让企业把数据变成策略，然后化为行动，继而找到新的解决方案，产生新价值的最大意义。

跨学科的新人才：数据科学家

另一方面，随着大数据应用的重要性益发为人重视，数字化经济也将会是接下来十数年间发展最快速的领域，伴随而来的人才需求，或者说人才短缺的问题也开始浮上台面。

例如，先前提到的数据科学家。对过去这些习惯默默无闻的工作，并且以冷静和理性著称的数据分析工作者而言，人们看待他们的眼光突然从过去的“数字怪咖”，变成了被《哈佛商业评论》认为 21 世纪“最性感”的职场人才。

这是数据科学从学术研究走向职业化的里程碑，如同我们前面提过的麦肯锡全球研究院研究报告，光是美国目前就需要 14 万到 19 万名的深度分析工作者，而商业智能公司 SiSense 的调查更发现，普通的数据分析师平均年薪 5.5 万美元，而副总裁级别的数据科学家年薪更可高达 13.2 万美元，而且因为人才短缺，在失业率居高不下的美国，该职缺 2013 年的年薪还在持续看涨中。



数据科学无疑是目前科技业增长最快的领域，目前美国好几所大学已有包括商业分析（Business Analytics）或是数据分析（Data Analytics）的硕士以上课程，专门培养前面所提到的“数据科学家”，而这门整合许多领域的专业技术，从华尔街的量化交易到网络上的广告定位，甚至是现实世界中的供应链优化都属于它的范畴。

那么，数据科学家的工作到底是什么？麻省理工科技创业杂志《Technology Review》访问了一位 Twitter 的数据科学家艾德温（Edwin Chen）。

为 Twitter 工作之前，艾德温在麻省理工学院学习的是纯数学和语言学，而他目前的工作内容包括构建计算机学习模型、进行数据可视化或是统计分析，以寻找更好的方式来理解 Twitter 上的使用者。

例如，我们可以从有关食物的 tweet 中得到些什么信息，如男人和女人吃的东西是一样的吗？旧金山和纽约的居民饮食习惯有何区别？用户发表的 tweet 和他们的饮食习惯有什么关联？或者，人们在伤心时会更倾向于吃垃圾食物吗？当然，最重要的是，这些海量文本经过计算机模型自动推断后，该如何将用户和广告商链接在一起。

简单来说，数据科学家的专长是“量化问题，然后解决问题”，他们的工作内容是 3 种东西的混合物：定量分析（使你了解数据），程序设计（使你可以处理数据），讲故事（让别人了解数据的含义）。

波士顿大学（Boston University）博士，MetaMarkets 共同创办人暨技术官德里斯科（Michael Driscoll）用另一种比较诗意的表述方式：数据科学家就像是目光如炬的探险家加上逻辑推论的大侦探，有如哥伦布和科伦坡的合体。

这样的描述还是无法让一般人知道，数据科学家到底应该具备哪些技能，我们也许可以从 Facebook 的职缺说明中找到答案。

在 Facebook 的职缺分类里，数据科学家被归类到"软件工程"中，但却比软件工程师更需要良好的沟通能力，以及重视产品的应用层面，而且必须是个 IT 通才。具体的招聘条件如下。

1. 相关技术领域的硕博士学位，或者是 4 年以上的相关应用经验。
2. 可以操作和分析从不同来源收集的复杂、大容量、高维度数据。
3. 热爱实证研究（Empirical Research），并善于用数据解答难题。
4. 能使用恰当的分析技能处理可用数据，灵活解答相关产品问题。
5. 与产品经理和工程师就数据分析结果进行沟通的能力。
6. 至少熟练掌握一种脚本语言（Python 或者 PHP）。
7. 有能力处理大规模数据，熟悉关系数据库和 SQL。
8. 能使用分布式计算工具如 Map/Reduce、Hadoop、Hive 等者优先。

Facebook 数据小组负责人马洛（Cameron Marlow）解释说，数据科学家必须肩负着从数以 TB 计的海量个人数据中，发现新商业价值的重任。他把自己的工作比喻为制造一个超高倍数望远镜，"就如同天文学改变这个世界对宇宙的理解一样，数据科学家开发的技术也可以改变这个世界对人类行为的科学理解。"

例如，人们怎么从社交网站上获得价值？一段时间后会有什么改变？这就像是一种"人类动力学"，找出如人际影响、关系强度、信息扩

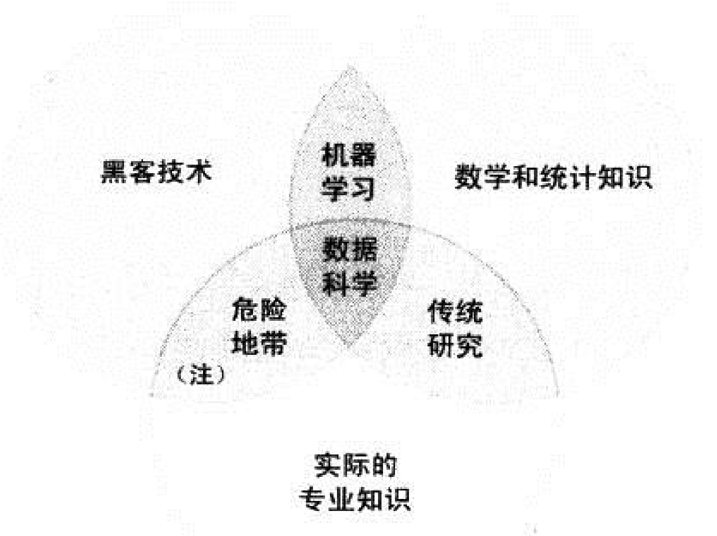


散和社会支持等各种因素的交互作用。

这些受过相关训练的数据科学家，在对于数字的好奇心和热情驱使下，将各种分析方式“混搭”，可以在已知的问题上找答案，在未知的问题中找知识，甚至在“不知为不知”（We don't know what we don't know）的环境下，找出这些过去不知道、不存在的模式，继而影响产品、流程和决策，这也是做为一个数据科学家最困难却也最具价值的地方。

现在，你可能对于数据科学家的工作内容已有粗略的认识，但到底什么样的人有能力做到这些呢？你可以把他们看成是一部分的数据黑客、一部分的数学专才、一部分的产业分析师、一部分的经营顾问，而且拥有用数据讲故事的特长。

图 14-1　数据科学家的角色



注：这种人有办法取得大量的数据，且因为扎实的专门背景，可以写出看似头头是道的分析。但因为他们没有统计背景，这些分析其实不准确。



以上这些特质，绝对不只是一般人所认为的，会写程序、会组装硬件的信息人才而已。因此，许多企业发现能够处理复杂数据的人才，可能都是有自然或社会科学领域的学习和工作背景，如生态学或是系统生物学。例如我们先前提到 Facebook 的马洛来自麻省理工学院媒体实验室，商业社交网站 LinkedIn 中的葛曼（Jonathan Goldman）是史丹福物理学博士，美国著名的支付服务软件公司 Intuit，旗下数据科学小组的负责人鲁曼利（George Roumeliotis）则是天文学博士毕业。

这些来自计算机科学、数学、经济学等有关于数据和计算密集型领域的跨学科人才，正在重组大数据分析的专业生态，尤其大数据本身具有横跨各产业的高度应用性，因此在企业组织内，跨领域的人才可能激荡出更多火花，例如 2012 年 7 月才上任的阿里巴巴集团首席数据官（CDO，Chief Data Officer）陆兆禧。

CDO 需具备的 5 种能力

1. 具备统计学、数学背景尤佳。
2. 洞悉网络产业和趋势发展。
3. 具备 IT 设备和技术选型的能力。
4. 商业营运的能力。
5. 管理和沟通的能力。

懂商业的经营新星：CDO

CDO 的重要性是从 2004 年开始，被誉为“数据挖掘之父”的乌萨玛·菲尔德（Usama Fayyad）担任雅虎的 CDO 之后逐渐在企业界发酵。相较于以研究为主的数据科学家，我们可以说，CDO 的定位是一位懂得商业运作的数据分析者，他必须从每天不断产生、繁复的大量数据中，解读什么是消费者想要，而且企业可以提供的服务，同时将两个端点串连成可以执行的行动方案。

真正的 CDO 需要集业务知识、IT 知识和经济学、统计学于一身，



不仅要关注的是系统架构中所承载的内容，更要担任企业决策和数据分析汇整的枢纽。他要熟悉包括面向服务的架构（SOA）、商业智能（BI）、大规模数据集成系统、数据存储交换机制以及数据库、可扩展标记语言（XML）、电子数据交换（EDI）等系统架构，也要深入了解企业的业务状况和所处的产业背景，更要清楚地知道企业的数据源、大小、结构等，才可以将数据资料与业务状况联合起来分析，提出相对应的市场和产品策略。

从企业的角度来说，CDO的知识架构应该分为信息技术（IT）与商业营销（Marketing）两个层面，才能为决策者提供“信息地图”和“数据仪表板”，让管理者可以清楚了解到数据带来的商业驱动力。

由于CDO的工作不仅仅是在技术层面，加上数据分析虽然独立于业务、IT和营销部门之外，但却同时需要和这些部门紧密相连，因此在组织管理上开始出现了不同的声音。有人认为数据管理隶属于IT部门的范畴，因为目前企业数据的主要来源还是各种IT系统，所以CDO应该与首席技术官（CTO，Chief Technology Officer）一起向首席信息官（CIO，Chief Information Officer）汇报。

也有人主张，在大数据时代的理想状态下，CDO必须整合数据价值与企业决策，所以应该至少是公司核心管理团队的5人之一（其他包括首席执行官CEO、首席运营官COO、首席财务官CFO和首席信息官CIO），直接向CEO汇报，阿里巴巴集团对于CDO的角色规划便是如此。

CDO的5种角色

1. 为数据发声（Be the voi- ce of the data）：支持和执行数据管理的策略、标准和制度化。
2. 权衡数据风险（Measure and manage data risk）：发展测量和预知数据风险的能力，并检查对于企业经营风险的影响。
3. 影响公司策略（Influence corporate strategy）：以有效的数据分析促使企业做出更具洞察力的决策。
4. 提高收入（Improve the top line）：利用数据增加营收、顾客认同、顾客保留率和商誉。
5. 提高利润（Improve the bottom line）：通过及时和正确的数据降低成本、提高生产力。



担任阿里巴巴集团CDO的陆兆禧可以直接向执行长马云汇报，无须经过CIO，这不仅引发中国内地的科技业一番震撼，CDO的权责也再次成为讨论焦点。另一个出人意表的事实是，陆兆禧并非DBA（数据库管理员）出身，且没有深厚的技术背景。

他在广州大学念的是酒店管理，从早上7点到晚上11点都在端盘子、擦桌子的酒店服务生做起，一路升到大堂经理、客房经理、餐厅经理。28岁时辞职，与朋友合伙经营一家网络通信公司，但并不成功。两年后加入阿里巴巴，毫无金融资历的他一手将支付宝打造为中国最大的支付系统，2008年担任淘宝网总裁后，让该网站的市场占比、人气都有了显著的提升，才开始被媒体所注意。

陆兆禧自己曾说，“做学生要读书，但不要迷信书；做总裁，要相信数据，但不要迷信数据，这样才不会盲目。”这代表的是，他相信数据所提供的现象，但他更重视数据分析背后代表的意义，而这也是阿里巴巴集团为何选择非技术出身的陆兆禧做为CDO，因为惟有深入了解业务的人才，才有能力真正的把数据变现金。

然而，是否每一家企业都需要把CDO纳入“C级俱乐部”（C-level Club）？事实上，每一家公司或组织都有自己的历史和特质，这不仅关系到数据管理的来源和类型，也左右了CDO在公司内部的必要性和功能，或许我们可以从下列“CDO函数”开始讨论。

1. 规模和历程（Size and footprint）：一家企业在该市场或国家的排名，包括市场规模、分公司或子公司数、员工和顾客数，原则上



规模越大、辖下公司和员工数目越多者，越有机会产生大量的数据需要管理和分析。

2. 产业（Industry）：一家企业的核心业务会产生大量且复杂的数据资料，例如金融服务业和制造业，抑或是与顾客互动需要倚赖高质量数据分析的产业，例如零售业或电子商务，又或者是对于数据需要高度监控的行业，例如金融业、医疗护理或是某些娱乐产业。
3. IT 组织（IT organization）：公司内部的 IT 部门是独立运作，还是用来服务其他部门？它的预算是集中管理，还是各个业务单位支付自己的技术需求？CIO 要报告的对象是 CEO 还是 CFO？换句话说，IT 部门被视为不可或缺的公司策略单位，或者它只是被视为一个成本中心，帮助日常业务顺利运作？如果答案都是前者，那么 CDO 的存在才有其意义。
4. 数据管理的成熟度（Maturity of data management）：包括数据质量、管理方式、主数据（master data）管理和数据安全，是否都有系统进行分析，并且成为企业决策的参考依据？是否有数据管理的政策和程序，而且采用产业标准？最重要的是，企业经营高层有决心支持数据管理计划。

综合上述，我们可以说规模大 × 适合的产业 × IT 部门的重要性 × 数据管理的成熟度 = 一个适合 CDO 施展才能的环境。然而，最好的计划也会因为没有利益关系人（stakeholders）的支持和参与而胎死腹中，企业



在设置 CDO 的时候应该特别注意以下几点。

1. 其他管理阶层支持与否（Total backing of the rest of the C-level）：这主要是确认 CDO 的权责领域。其他高层主管的理解、支持和积极参与，以及对于 CDO 这个角色和职责的认同都相当重要，因为这关系到 CDO 往后如何整合各单位的数据或人力资源。
2. CDO 应肩负大型的业务问题（Piggyback on large business problems）：在大多数的组织中，只是利用数据说话势必难以服众，除非 CDO 可以解决并直接响应一些迫在眉睫的大型业务问题。例如，金融业如何利用数据分析应对最近加强查核的监管法令，或是如何帮助医疗业应对法令规定建立标准化的电子病历等。也就是说，CDO 的功能在于帮助企业的"核心业务"（heart of the business）解决问题或应对挑战，这样，较可能避免数据小组在发展的婴儿期就面临夭折。
3. 识别有特定综合技能的 CDO（Recognize the specific combination of skills of a CDO）：如前所述，一个好的 CDO 必须同时平衡技术技能、商业知识和人际沟通能力，可以担任数据小组的领航者，也必须有深厚的商业市场专业，更要有能力解决办公室内的政治问题，才能够使得项目顺利推行。
4. 对于 CDO 的责任应该明确定义（Clearly define respon-sibilities for the CDO）：各个企业对于 CDO 的角色和职责必须要明确的规范，



例如，最早开始聘用CDO的华尔街，就很清楚地知道CDO的职责是关注金融风险管理和符合法规；其他产业的CDO主要职责可能在于市场分析、供应链数据可视化（supply chain visibility）或顾客管理。如此一来，企业内的其他单位也会对CDO的权责，和需要提供或配合的协助较为清楚。

5. 赋予CDO执行权（Empower the CDO with execution autho-rity）：没有任何权力和资源来执行数据管理计划，CDO就不可能真正影响到业务，因此除了配备IT和业务资源之外，CDO也要有执行企业策略或项目的权力，才能真正发挥数据管理和分析的作用。

是童话还是事实？是文明还是老大哥？

的确，不论是数据科学家或是CDO，都是这个“与数俱进”的时代里新登场的重要角色。但是我们也要告诉你的是，组织变革是相当困难的！

这就像是童话故事的开头通常是“很久很久以前……”，结束时通常是“从此，他们过着幸福快乐的日子”一样，没有人去解释为什么童话故事中的王子和公主，结婚后不会争吵、永远和睦。

也就是说，大数据可以转化公司业务、改变企业经营者的思维，但却无法解释或说明“如何”改变公司内部的组织，甚至根据我们的经验发现，当一家企业的问题可以利用数据分析来改变时，它最急迫需要的



却不是数据分析，而是人们在组织架构图中的角色，以及系统设计过程中的变化。

让我们假设数据分析的确揭示了业务改进的机会，无论是在降低成本还是增加收入方面。然而真正要做到，必须双管齐下，改变组织架构中的流程和角色，同时追踪这些变化的结果，建立一套双向反馈的管理制度。另一个大多数企业会面临的挑战是，在进行数据管理和分析时，是否每一个人都在同一条船上？当有参与者不一起玩，在这个升级过程中是否有惩罚性措施？

我们回头去看看一个经典案例：在 CompStat 的帮助下，纽约市长朱利安尼成功减少了犯罪率。之前，我们提过在每周一次的会议中，纽约市警察局会召集旗下 5 个区指挥官讨论问题、制定战略和战术来减少犯罪。但是你可能不知道的是，头一年他们讨论最多的是如何重组人事结构、建立执行流程以及设置问责制，而且根据 CompStat 追踪的结果，这些战略和战术的成功或失败，分区指挥官都会被追究责任，以确保这些变革得以真正被执行且受到管控。

所以，当有人声称大数据将改变你的业务，请记住，它的确带来了转型的潜力，而不是实际上的转型。因为，这些还需要加上承诺和努力工作，否则再怎么精密的分析都像是你原来所想，“从此过着幸福快乐的生活”只不过是童话而已。

所有事情都有一体两面，大数据对企业来说，是童话还是事实？要看企业是否有管理转型的决心。另一项值得反思的现象是，大数据对这个世界来说，是文明还是老大哥（Big brother）？这得要看你我有没有自



律的决心。

“不论是睡着还是醒着，在工作还是在吃饭，在室内还是在户外，在澡盆里还是在床上——没有躲避的地方。除了你脑壳里的几立方厘米以外，没有东西是属于你自己的。”

这是乔治·奥韦尔（George Orwell）著名的小说《一九八四》里的情节。这部刻画出人们对于专制集权的反思之作，描述在大洋国里的居民，党领袖“老大哥”利用电幕（telescreen，一种双向电视）来监控人民的表达、交流和生活，任何一个对党不敬的言论，甚至是表情，都可能被报告，然后是教育、逮捕、改造、清洗。

在那个钳制思想、生活贫乏的世界里，完全没有隐私和自由可言。而这种通过科技形成的监控也让人们开始对于“大数据是否等于老大哥”产生争议，认为科技的进步和政府或企业权力的扩张，将会对个人隐私造成潜在的威胁。

当然，这种担心并非空穴来风。当美国大卖场 Target 比父母还要早知道女儿怀孕的消息（利用购买数据推测），银行比妻子们更清楚先生光顾了特种场所[1]时，个人数据的隐私性问题就更被大众所注意。

例如，美国前总统布什（George W. Bush）在 2002 年，下令美国国家安全局（NSA）监听美国公民对国外的通话，以便及早发现恐怖分子，

[1] 美国某银行装设一台新的 ATM 提款机，但接下来几周的数据记录却非常奇怪，因为每天的午夜 12 点到 2 点之间，都有大量的款项被提取。该银行担心这种现象涉及诈骗等违规操作，于是雇用侦探进行监控和调查。结果发现，之所以有很多顾客选择在午夜提款，是因为这台提款机就设在一家色情俱乐部旁，顾客不想在信用卡上留下消费记录，所以改为提取现金。后来，当地的报纸还以“XX 银行知道昨晚谁光顾了妓院”为题报导此事。



因而建立了一个全美最大的个人信息数据库，收集几千万人的地址、电话和其他数据。这项行动在2006年被《今日美国》踢爆后，人们才发现这些侵犯隐私的政府违法作为，背后都有民间电信、电话公司的配合。报导中甚至还指称联邦调查局（FBI）希望在包括Facebook和Twitter等社交网站上建立可监控的“后门”，这代表的是美国政府正走向奥韦尔式（Orwellian）的状态。

此举引起群众哗然，甚至美国有些律师主动推行修法，希望把“个人数字信息”纳入美国宪法《第4修正案》。法案原文是：“人人具有保障人身、住所、文件及财物的安全，不受无理的搜索和拘捕的权利；此项权利，不得侵犯；除非有可成立的理由，加上宣誓或誓愿保证，并具体指明必须搜索的地点，必须拘捕的人，或必须扣押的物品，否则一概不得颁发搜捕状。”

由于目前无明文规定“隐私权”为保障的范畴，因此相关人士希望纳入“人民数字信息和通信的隐私权是不可剥夺的”（The people's right to privacy in their data and communications shall not be abridged.）等文字内容。

不可否认地，政府或企业运用大数据是恰如其分，还是稍一不慎便成为老大哥的情况，经常只有一步之差，但也不可因噎废食，忽略了大数据分析可能为医疗、公共安全等领域带来的社会益处，以及在商业世界中企业可以为人们提供更适合选择的市场优点。

重点是，在运用大数据的同时，政府和企业的目的性。我们没有办法针对每一个政府、每一家企业的情况，制定一套共同的标准，帮助它们顾全更多人的利益，同时保护更多人的隐私，但必须强调的是，不论



是政府或是企业，都必须确保他们不要跨越自定义的界限，造成隐私入侵的后果。

科技本身是中立（neutral）的，如同《圣经》所说：上帝叫日头照好人，也照歹人；降雨给义人，也给不义的人。科技的演进会让好事发生，也会让坏事发生，但我们相信，在人类迈向更文明、更美好的过程中，唯一的纪律，就是持续自律（the only discipline that lasts is self-discipline），我们也衷心期盼在人类文明的演进过程中，良善的一面可以获得多一点帮助，让大数据分析协助产业建立新价值、找到新方向，创造更多的就业新机会，让这个世界达到中国人所说的“大道之行也，天下为公”的境界。

（本书由方沛晶、施祖琪采访整理）